A strategy of Mapping Polish WordNet on Princeton WordNet

Ewa Rudnicka*, Marek Maziarz*, Maciej Piasecki*, Stanisław Szpakowicz**
*) Wrocław University of Technology, **) University of Ottawa

Abstract

We present a strategy, and the early results, of mapping plWordNet – the largest Polish wordnet – onto Princeton WordNet. The fundamental structural premise of plWordNet differs from that of most existing wordnets: its basic building block is the lexical unit rather than the synset. The addition of new material is consistently informed by semantic relations, and by a range of analyses of large corpora. The relative independence of plWordNet of the design principles of WordNet and WordNet’s focus on synsets pose a challenge to any mapping between the two lexical networks. Mapping between WordNet and “national” wordnets tends to rely on a translation-based transfer-and-merge method. We have developed a set of inter-lingual relations and a mapping procedure which allows us to link successfully two structurally distinct systems.

1 Introduction

We present a strategy and the preliminary results of the mapping of Polish WordNet [plWordNet] onto Princeton WordNet [PWN] (Fellbaum 1998). There have been many attempts to build such mappings for wordnets for both European and non-European languages, including EuroWordNet [EWN] (Vossen 1998, Vossen 2002), MultiWordNet (Bentivogli, Pianta, Pianesi 2000; Bentivogli, Pianta 2000), AsianWordNet (Robkop et al. 2010) and IndoWordNet (Sinha, Reddy, Bhattacharyya 2006, Bhattacharyya 2010). Those projects usually took advantage of the transfer-and-merge method developed during the construction of EWN. The method’s main stage is the translation of most of PWN’s structure and content into the target language, so in effect a mapping from one side of such a “bilingual wordnet” to the other side is quite obvious. Unlike those wordnet systems, plWordNet’s design and construction are independent of EWN or PWN, though inevitably substantially influenced by both. A unique corpus-based method was employed (Maziarz, Piasecki, Szpakowicz 2012, Piasecki, Szpakowicz, Broda 2009). Synsets in plWordNet are merely groups of similarly interconnected lexical units, and it is the lexical unit that is the basic element of the network. We aim, then, to link two largely independent lexical systems. An inter-lingual mapping procedure connects plWordNet synsets with PWN synsets via an ordered set of inter-lingual semantic relations. Mapping is manual, but it is very strongly supported by automatic prompting. Given that nouns are by far the most numerous class in PWN and in plWordNet, we have decided to begin testing our procedure in practice by the mapping of nouns in specific domains. The domains include family relations, names of people and places, food and drinks, clothes, plants and animals, basic economic vocabulary, and units of time.

2 plWordNet and Princeton WordNet

Princeton WordNet and plWordNet have been built on somewhat different philosophical, theoretical and methodological grounds. We review the principles of the construction of both wordnets, and name problems which may arise in the process of mapping. The clearest differences are to do with the concepts of synonymy and synset.
According to Miller (1998), the synset is WordNet’s basic building block. A synset represents a “lexicalized concept”. Synonymy is considered the fundamental semantic relation in PWN. Any decision on synonymy is made in view of linguistic intuition supported by dictionary queries. Synsets – sets of synonyms – are used to construct hyponymy hierarchies (Miller 1998: 23-28).

The philosophy of plWordNet is different (Piasecki, Szpakowicz, Broda 2009). It is the lexical unit [LU] that resides at the center. Lexico-semantic relations are defined for LUs. LUs which share targets of certain wordnet relations, such as hypernymy or meronymy, are grouped into derived structures to which we refer as synsets. Those wordnet relations are copied from LUs to synsets. Unlike in PWN, then, we start with LUs and assign to them proper lexico-semantic relations. In the end, we obtain sets of LUs indistinguishable from each other in terms of their connectivity in the network.

In PWN, relations such as hyponymy or meronymy are conceptual relations (Miller 1998: 25-8, 37-9), while in plWordNet they are lexico-semantic relations. This structural difference results in the average synset smaller in plWordNet 1.6 (nouns: 1.37 LUs per synset, verbs: 1.47, adjectives: 1.38) than in PWN 3.0 (nouns: 1.78, verbs: 1.81, adjectives: 1.65). These numbers could be interpreted as follows: synonymy in plWordNet is somehow more restrictive than in PWN.

In effect, a mapping between PWN and plWordNet is not easily done. That is because it is not obvious what to link: lexical units (as in an ordinary bilingual dictionary) or synsets (the approach likely to be accepted by most of wordnet designers).

3 Existing multi-lingual wordnets

There have been a number of projects to link Princeton WordNet with other wordnets, and to build new wordnets closely aligned with PWN. The EuroWordNet project was launched to build a multi-lingual lexical database consisting of inter-linked “national”, mono-lingual wordnets for Dutch, Spanish, Italian, French, German, Czech and Estonian, and to link this multi-lingual database to the original PWN. Thus, the structure of EWN has from the very beginning been a very close reflection of the structure of the original. Each individual “national” wordnet was based on the same theoretical and methodological principles as PWN.

Now, to map EWN wordnets onto PWN, an inventory of inter-lingual relations was needed. The proposed list largely mirrors the intra-lingual relations in PWN – and nowadays in many other mono-lingual wordnets. The essential relations include inter-lingual synonymy and near-synonymy, inter-lingual hyponymy and hypernymy, and inter-lingual meronymy and holonymy. We will now discussed these relations.

The key inter-lingual relation is inter-lingual synonymy, defined as “one-to-one mapping between synsets” and based on the assumption that “one source synset is linked to one and only one target synset” (Vossen 2002). This relation entails the identity of the denotations of the word-sense pairs in the source and target language. So, all lexemes with the given sense are grouped into one synset both in the source- and the target-language wordnet. The near-synonymy relation is postulated for the less clear cases of inter-lingual synonymy, when more than one target synset is to be linked with the source synset, or more than one source synset with one target synset. Inter-lingual hypernymy holds when the source synset has a more specific meaning than the

---

1For example, PWN groups common and scientific names for a cat in one synset {domestic cat, house cat, Felis domesticus, Felis catus}, while in plWordNet they are split into two synsets {kot ‘cat, domestic cat’} and {kot domowy ‘a Polish taxonomic term for Felis domesticus’} (there are no Latin names in plWordNet).
target synset, hyponymy – when the source synset can only be linked to more specific target synsets.

**MultiWordNet** groups several Romance languages. Italian is the core wordnet of the project, while Spanish, Portuguese, Romanian and Latin wordnets are compatible with it. Italian WordNet is built up from PWN by a semi-automatic acquisition method. The creation of Italian synsets is based on a monolingual Italian dictionary and then linked to the corresponding PWN synset via a relation of correspondence whenever it is possible (Bentivogli, Pianta, Pianesi 2000). According to the Authors, the most important difficulty are lexical gaps – when there is no equivalent either in the Italian or in the English wordnet (Bentivogli, Pianta 2000) – and denotation overlap – when “a translation equivalent exists, but it is more general or more specific” (multiwordnet.fbk.eu/english/whatin.php). In the latter case, a special NEAREST relation is applied (“usually a hyponym or a hypernym”, *ibid.*).

**IndoWordNet** was built by the expansion method from HindiWordNet (Bhattacharyya 2010). It groups 16 of India’s 22 languages. It is also linked to the English synsets in PWN. According to (Bhattacharyya 2010), “13693 synsets of Hindi have been linked with English. Efforts are also on to automatise this process of linkage. (...) it was emphasized again and again to the lexicographers never to translate the words in the Hindi synset, but to understand the meaning expressed by the synset and its attached gloss and example sentence and then to put down the words in frequency order the words of the language.” (Sinha, Reddy, Bhattacharyya 2006) present a strategy for relation-borrowing between Hindi and Marathi:

“When the sense is present in both Hindi and Marathi: The relations are established for that sense.

When the sense is in Hindi but not in Marathi: The relations will not get established for that sense.

When the sense is not in Hindi but in Marathi: The relations for that sense in Marathi have to be established manually.”

### 4 The typology of inter-lingual relations

This section presents a typology of inter-lingual relations employed in the process of mapping *pl*WordNet onto Princeton WordNet. The relation set includes inter-lingual synonymy and near-synonymy, inter-lingual hyponymy and hypernymy, and inter-lingual meronymy and holonymy. We have made one central decision: once the first possible link has been created, the remaining relations are no longer searched for and applied. The order and the conditions of the application of the specific relations are motivated by their position in a ranking, which we also describe in this section.

**Inter-lingual synonymy** is defined as a correspondence between the senses (the identity of the denotation) and the position of the source and target synsets in the wordnet structure. It will be applied in a few cases. The easiest variant is the situation when both the sense and the position in the wordnet structure of the source and target synsets are the same:

\[ \text{(Example 1)} \quad \{\text{wino 1}\} \leftarrow \text{I}-\text{synonymy} \rightarrow \{\text{wine 1}\} \]

\[ \{\text{wino 1}\} \leftarrow \text{hypo} \rightarrow \{\text{alkohol 1}\} \quad \{\text{wine 1}\} \leftarrow \text{hypo} \rightarrow \{\text{alcohol 1}\} \]

\[ \{\text{wino 1}\} \leftarrow \text{hyper} \rightarrow \{\text{białe wino 1}\} \quad \{\text{wine 1}\} \leftarrow \text{hyper} \rightarrow \{\text{white wine 1}\} \]

\[ \text{The inter-lingual nature of a relation is signalled in this paper by the prefix } I-\text{.} \]
The literal translation of *białe wino* is ‘white wine’, *alkohol* is an exact Polish equivalent of *alcohol*, and many other Polish and English hyponyms and hypernyms of *wino* and *wine* are directly comparable. So, we can say that the structures of *plWordNet* and *PWN* are the same in the vicinity of these LUs. Knowing the denotation of *wino 1* and *wine 1* (which is the same) leads us to the conclusion that synsets containing the two LUs must be in the *I*-synonymy relation.

It is a slightly less obvious situation when the immediate hypernym cannot be linked by *I*-synonymy, but the one located one or more levels above it can.

(Example 2)

{**babcia 2**, babunia 1, babula 2} ←*I*-synonymy→ {**grandma 1**, grandmother 1, granny 1}

{**babcia 2**} —hypo→ {przodkini 1}       {**grandma 1**} —hypo→ {grandparent 1}

{**babcia 2**} —meronymy→ {dziadkowie 1}     {**grandparent 1**} —indirect hypo→ {ancestor 1}

{**grandma 1**} —hyper→ {nan 1}

*babcia* is connected with *przodkini* ‘female ancestor’ via hypernymy and with *dziadkowie* ‘grandparents’ via holonymy, while *grandma* has links to *grandparent* (a hypernym) and to *nan* (a hyponym). The closest equivalent of Polish *przodkini* is English *ancestor*, which is an indirect hypernym of *grandma* (through *grandparent* and *forebear*). It is obvious that *babcia* means the same as *grandma* (the denotation is the same), although the network structures differ. We would say that analogous wordnet structures suffice to establish *I*-synonymy between the two synsets.

*I*-synonymy is in fact an identity relation of the *I*-synonyms’ *denotata*. **Inter-lingual hyponymy** should be described in terms of denotation set inclusion: a hyponym refers to an object which is included in a denotation set of a hypernym. The relation is more distant than between *I*-synonyms, so it is the second in our ranking of inter-lingual relations. It is defined in the following way: the relation links a kind of an entity (named by a source-language synset) to its proper kind term (expressed by a target language synset). Consider an example:

(Example 3)  {**wycieczka 2**} —*I*-hypo→ {**party 3**, company 7}

{**wycieczka 2**} —hypo→ {ludzie 1}       {**party 3**} —hyper→ {set 5}

{**wycieczka 2**} —holonymy→ {wycieczkowicz 1}       {**party 3**} —hyper→ {rescue party 1}

*Wycieczka* is a group of people (‘party’) assembled temporarily for the participation in a trip; it consists of *excursionists* (Pol. *wycieczkowicz*), and it groups *people* (Pol. *ludzie*). There is no equivalent of the noun *wycieczka* in English. Instead, it must be linked to the closest noun which, in this case, is the hypernym *party* (company).

It may happen that a Polish or English word with no *I*-synonym is linked to a word with a narrower extension:

(Example 4)  {**cukiernia 2**, ciastkarnia 1} ←*I*-hyper→ {**patisserie 1**}

{**cukiernia 2**} —hypo→ {sklep 1}       {**patisserie 1**} —hypo→ {bakery 1}

{**cukiernia 2**} —hyper→ {cukierenka 1}       {**patisserie 1**} has no hyponyms

Polish *cukiernia* is a kind of shop (sklep) or bakery in which pastry is sold. It does not have a direct synonym in English. It seems that the closest equivalent is *patisserie* but it is a term borrowed from French, somewhat less broad.

---

3 For simplicity we show only the most illustrative existing links between synsets {wino 1} and {wine 1}.
When neither inter-lingual synonymy nor hyponymy can be applied, **inter-lingual meronymy** is the third-ranked option in our linking procedure. We have assumed that direct links to *I*-synonyms or links to *genera proxima* will be more useful than connections referring to parts and wholes. Here is an example of such linkage:

(Example 5) \{wyleżysko 1\} —*I*-meronymy→ \{hillside 1\}

\{wyleżysko 1\} —hypo→ \{miejsce 1\} ‘place’

\{wyleżysko 1\} —meronymy:place→ \{zbocze 1, stok 2\} ‘hillside’

*Wyleżysko* is a place where snow has lain for a long time. It has no equivalent in PWN. Linking it with *I*-hyponymy is not right, because the closest PWN hypernym is \{topographic point 1, place 1, spot 1\} ‘a point located with respect to surface features of some region’ – still too remote a connection. Thus the LU is connected to *hillside* ‘the side or slope of a hill’ by *I*-meronymy, analogously to the *meronymy* link between *wyleżysko* and *zbocze/stok* in *plWordNet*.

An inverse relation is **inter-lingual holonymy**. A *plWordNet* synset denotes a whole whose part is designated by a PWN synset, as in the following example:

(Example 6)

\{zagranica 1, obczyzna 1, obce terytorium 1\} —*I*-holonymy→ \{foreign country 1\}

\{zagranica 1\} —hypo→ \{strefa 2\} {foreign country 1} —hypo→ \{state 4\}

\{zagranica 1\} —meronymy→ \{świat 2\}

Polish *zagranica* ‘countries abroad’ has no *I*-synonym, no *I*-hyponym nor inter-lingual hypernym. Indeed, *zagranica* is a kind of a geographical area (*strefa*) which is outside state borders, and it is a part of our world (*świat*). It is natural to see a whole consisting of a foreign country in a set of foreign countries. That is why an *I*-holonymy relation has been applied.

**Inter-lingual near-synonymy** is proposed as a last resort, when none of the previous relations applies. It will be used for cases of the semantic or structural overlap between the senses and networks of relations of the source and target synsets. An example of such a situation is the case when there are two candidate target synsets for one source synset and both of them share part of the meaning or structure of the source synset. As for the structure, this boils down to the overlap between the sets of hyponyms of the two candidate synsets and the source synset. Here is an example of such overlap:

(Example 7) \{pracownia 1\} ←*I*-near synonymy→ \{workshop 1\}

\{pracownia 1\} —hypo→ \{pomieszczenie 3\} \{workshop 1\} —hypo→ \{workplace 1\}

\{pracownia 1\} —hyper→ szwalnia 1 \{workshop 1\} —hyper→ \{bookbindery 1\}

\{warsztat 2\} ←*I*-near synonymy→ \{workshop 1\}

\{warsztat 2\} —hypo→ \{pomieszczenie 3\}, \{warsztat 2\} —hyper→ \{introligatoria 1\}

It is impossible to attribute any of the relations considered thus far to *pracownia* or to *warsztat*. *Pracownia* is a kind of room or building (*pomieszczenie*) in which a craftsman or artisan works and goods or artworks are produced or repaired. A kind of *pracownia* is *szwalnia* (sewing room). *Warsztat* could be perceived in some cases as a kind of *pracownia*, but not all hyponyms of *warsztat* are hyponyms of *pracownia*. For example, a place where books are bound by a bookbinder (*introligatornia*) could be both *warsztat* and *pracownia*, while a repair shop for automotive vehicles (garage, Pol. *warsztat samochodowy*) could be described only as a kind of
Both nouns (warsztat i pracownia) roughly cover the range of workshop, but some subtle differences remain.

A relation additional to all those previously mentioned is I-inter-register synonymy. It links synsets with the same denotation but different stylistic register. It is applied to Polish synsets otherwise linked via bzn – plWordNet’s inter-register synonymy – to plWordNet synsets which already have their I-synonyms in PWN. Thus, their bzn synonyms in plWordNet are linked via I-inter-register synonymy to their I-synonyms in PWN, as illustrated below:

(Example 8) \{bachor 1, ...\} ‘brat’ ← bzn → {dziecko 1, ...} ‘child’
{dziecko 1, ...} ← I-synonymy → {child 1, kid 1, ...}
{bachor 1, ...} ← I-inter-register synonymy → {child 1, kid 1, ...}

5 Mapping strategies

We now briefly describe the mapping strategies we have been applying in the process of mapping plWordNet and PWN, and the motivation behind them. Here are the key components of our mapping procedure: recognize the sense of the source synset, search for the equivalent target synset and link the source and target synset by one of the I-relations presented in section 4.

Our first order of business is the mapping from plWordNet to PWN, so plWordNet synsets will be our source synsets. The first step in our manual mapping procedure is the proper identification of the sense of the source synset. Unlike the majority of PWN synsets, very few plWordNet synsets have glosses, although the considerably more frequent comments partly make up for the absence of glosses. Still, plWordNet is largely relation-based, so the key (sense) denominator will be the position of the given set of synonymous LUs in the overall wordnet structure. Nevertheless, the plWordNet editor begins with reading all lexical units in the synset, plus the glosses or comments if there are any. Take Example 6: \{zagranica 1, obczyzna 1, obce terytorium 1\} (countries abroad, foreign lands, foreign territory). Next, the editor considers the wordnet structure – looks at the immediate hypernyms and hyponyms as well as meronyms and holonyms, if there are any. There are strefa and świat (zone, world) in Example 6. In case of doubts or difficulties with determining the sense of the synset, the editor is advised to consider the hypernyms (or other relations) located one or more level above the direct hypernym.

Once the sense of the analysed synset has been established (‘area located beyond the borders of a given country’), the editor can move to the next stage: seek the equivalent target synset in PWN. The first step is to check the automatic prompt if there is any.

Definitions of relations refer to a good deal of linguistic knowledge, and that makes an automated decision procedure rather hard to implement. Bilingual dictionaries describe relations among lemmas in two languages, but do so implicitly, and in any case deep dictionary search can be time-consuming. It might be done in advance and the structures of two wordnets utilised in order to establish an initial draft mapping, which needs further verification and correction by linguists. We implemented an automated mapping algorithm described in (Daudé, Padró, Rigau 2003, Daudé, Padró, Rigau 2000). The algorithm is based on the Relaxation Labelling scheme. An initial mapping is generated from bilingual dictionaries and then iteratively corrected using a set of constraints on the structure of both wordnets and the links established so far. The algorithm was applied to the data in several bilingual Polish-English dictionaries. Three are publicly available: (Piotrowski and Saloni 2002), (Polish Wiktionary) and inter-lingual links in (Polish
Wikipedia); one is proprietary. The resulting dictionary included 597108 bilingual lemma pairs, both one-word and multiword; there were 501064 one-to-one and 96044 one-to-many translations. Evaluation on 1000 randomly selected synsets with inter-lingual links (manually added) showed that 90% automatically created links were supported by the linguists’ decisions, including 50.2% I-synonymy links. For relations different than I-synonymy, we also counted indirect links as correct, including non-direct hypo/hypernymy links.

The algorithm returns from none to several weighted mapping suggestions per source synset. After a manual analysis of the results, we have decided to utilise only those cases in which exactly one suggestion was returned. For plWordNet’s nominal part and the combined dictionary we found 27000 such synsets. The coverage of 27000 links seems to be large, but the suggested links were scattered across plWordNet’s hypernymy structure. Many of them were also related to such simple cases as names of animals, plants or substances, while we asked the linguists to focus first on selected hypernymy subhierarchies. Only automatically generated suggestions for synsets located in the selected subhierarchies could be utilised. For the presentation of the automatic mapping we used WordnetLoom (Piasecki et al., 2010), a visual, graph-based wordnet editor available on a GPL licence. Each suggested link is visually presented on the screen as an instance of an auxiliary inter-lingual relation called potential equivalent. Such links associate hypernymy structures in two wordnets, and selected subgraphs are presented on the same screen in parallel. The linguists can compare them and directly operate on them, adding new inter-lingual relations as well as changing the plWordNet structure.

If there is no automatic prompt, the next step is to refer to the editor’s language intuitions and select one or two candidates of target language lexical units which share the sense of the source language synset (‘foreign country’). Next, these candidate lexical units are found in PWN and their synsets are analysed with respect to their sense and position in the wordnet structure (hypernym state). Special attention must be paid to their immediate hypernym(s) and hyponyms (as well as to other relations if there are any), since these are going to be juxtaposed with the equivalent relations of the target synset. The linguist has to check if there already exist, or if there are likely to be postulated, inter-lingual synonymy links between any of the immediate relations of the source and the target synset. When such links exist or are likely to be established for the majority of the immediate relations, and the gloss of the target synset also matches the sense of the source synset, the inter-lingual synonymy is granted between the two synsets in question. If such links do not exist or are not likely to be established, the next candidate is verified.

When the editor decides that there are no more candidates to test, the previously chosen candidates are checked for their potential of linking via other relations. The remaining relations are tested in the following order: hyponymy, hypernymy, meronymy, holonymy and near-synonymy. In Example 4, we test the possibility of linking our source synset with the target synset {world 4, earth 9, Earth 1, globe 1} —hyper→ {terrestial planet 1}; —meronym→ {solar system} via I-meronymy, because this synset is a potential synonym of {świat 2}, a meronym of our source synset. Still, our source synset {zagranica 1, obczyzna 1, obce terytorium 1} belongs to the domain of political organization, while the target synset belongs to the domain of geography, so the link must be dismissed. Next, we check the potential for linking of the candidate target synset ‘{foreign country 1} —hyper→ {state 4}’ and decide that our source synset can be linked to this target synset via I-holonymy.

---

4 It was kindly made available to us free of charge by TiP Ltd. (www.tip.net.pl/).
6 Results and mapping dilemmas

Our mapping project began in March 2012. As of 2012-08-28, 18280 \textit{I}-relation instances have been introduced into \textit{plWordNet}; see Table 1.

<table>
<thead>
<tr>
<th>\textit{I-synonymy}</th>
<th>\textit{I-hyponymy}</th>
<th>\textit{I-hypernymy}</th>
<th>\textit{I-meronymy}</th>
<th>\textit{I-holonymy}</th>
<th>\textit{I-near-synonymy}</th>
<th>\textit{I-inter-register synonymy}</th>
</tr>
</thead>
<tbody>
<tr>
<td>7827</td>
<td>7468</td>
<td>1640</td>
<td>490</td>
<td>231</td>
<td>479</td>
<td>145</td>
</tr>
</tbody>
</table>

Table 1. Number of inter-lingual relation instances

The frequency of specific relations agrees with the proposed ranking, based on our intuitions concerning meaning closeness and the identity and inclusion of \textit{denotata} sets. Surprisingly, \textit{I-hyponymy} (\textit{I-hypernymy}) accounts for more than one third of all inter-lingual relations. This suggests that the structures of \textit{plWordNet} and PWN differ non-trivially.

In the course of mapping, we have faced dilemmas resulting both from the differences in the conceptual and lexico-grammatical structure of English and Polish, and from different methodological assumptions which underlie the construction of \textit{plWordNet} and PWN. We will discuss all types of these dilemmas with an eye to their possible solutions.

6.1 Lexico-grammatical differences

6.1.1 Lexicalisation

The obvious trouble with the mapping between \textit{plWordNet} and PWN is the existence of lexical gaps: concepts either not lexicalised in one of the languages or non-existent in its extra-linguistic reality and conceptual structure (cultural gaps). An example of the former is the English word \textit{chantry} meaning “a chapel endowed for singing Masses for the soul of the donor” (adopted from PWN’s definition of \{chantry 2\}). The concept is not lexicalised in Polish, though it exists in the extra-linguistic reality, so \{chantry 2\} is linked to its closest Polish equivalent \{kaplica wotywna 1\} via \textit{I-near synonymy} which signals partial correspondence in meaning or structure:

\[
\begin{align*}
\text{(Example 8)} & \quad \text{\{chantry 2\}} \rightarrow \text{\{hypo\}} \rightarrow \text{\{chapel 1\}} \rightarrow \text{\{hypo\}} \rightarrow \text{\{place of worship 1\}} \\
\text{\{kaplica wotywna 1\}} & \rightarrow \text{\{hypo\}} \rightarrow \text{\{miejsce kultu 1\}} \\
\text{\{chantry 2\}} & \leftarrow \text{\{I-near-synonymy\}} \rightarrow \text{\{kaplica wotywna 1\}} \\
\text{\{place of worship 1\}} & \rightarrow \text{\{I-hypo\}} \rightarrow \text{\{miejsce kultu 1\}}
\end{align*}
\]

Cultural gaps can be the names of occupations or administrative functions never present in the other language’s reality, so not lexicalised. An apt example is \{kaowiec 1\}, a Polish term denoting an institution’s employee responsible for the organization of cultural and recreational events in the Communist times. It is linked to the PWN synset \{organizer 1 ...\} meaning “a person who brings order and organization to an enterprise” via the \textit{I-hyponymy} relation, which is the most often used relation in such cases:

\[
\begin{align*}
\text{(Example 9)} & \quad \text{\{kaowiec 1\}} \rightarrow \text{\{hypo\}} \rightarrow \text{\{pracownik oświaty 1\}}; \text{\{organizer 1\}} \\
\text{\{organizer 1 \}} & \rightarrow \text{\{I-hyper\}} \rightarrow \text{\{organizer 1 ...\}} \\
\text{\{kaowiec 1\}} & \rightarrow \text{\{I-hypo\}} \rightarrow \text{\{organizer 1 ...\}}
\end{align*}
\]
The last type of lexical gaps involves mismatches resulting from different structuring of information, as in the case of English and Polish family relation hierarchies. Polish lexicalizes the distinction between the brother of one’s father (straż or stryje) and mother (wuja or wujek), although the former term is marked and slowly becomes obsolete. Both terms are present in plWordNet. The unmarked term {wuja 2} is linked to its English equivalent {uncle 1} via the I-synonymy relation, while the marked term {straż 1} is linked to {uncle 1} via I-hyponymy:

\[
\text{(Example 10)} \quad \{\text{straż 1}\} \rightarrow \text{hypo} \rightarrow \{\text{wuja 2}\}
\]

\[
\{\text{wuja 2}\} \rightarrow \text{I-near-synonymy} \rightarrow \{\text{uncle 1}\}
\]

\[
\{\text{straż 1}\} \rightarrow \text{I-hypo} \rightarrow \{\text{uncle 1}\}
\]

Still, it is possible to express the contrast in English using the premodifying adjectives paternal and maternal, except that the resulting nominal phrases paternal uncle and maternal uncle are not lexical units in PWN.

It is important to distinguish all these gaps from dictionary-content gaps due to differences in sources or methodology of building the two wordnets. Clearly, our most preferred I-synonymy relation cannot be used in either instance. Still, most of these cases can be handled by the I-hyponymy/hypernymy relation which we treat as the second option. Occasionally, we resort to I-meronymy/holonymy relations and I-near-synonymy relation.

6.1.2 Feminine forms

Another type of dilemma is to do with the divergent degree of gender lexicalisation in English and Polish. Feminine nominal forms in Polish are frequent, while most of English nouns are not marked for gender. An example is the English word cousin and Polish kuzyn ‘cousin_masc’ and kuzynka ‘cousin_fem’. The most natural strategy to adopt here is again to resort to I-hyponymy, making the English {cousin 1} the hypernym of both Polish {kuzyn 1} and {kuzynka 1}, which can easily be construed as two sub-types of a more general concept. Interestingly, there are also mixed English synsets consisting of feminine and masculine forms (and sometimes also unmarked forms), as in {bondswoman 1, bondsman 2} or {chairman 1, chairwoman 1, chairperson 1}. I-hyponymy links such synsets to the corresponding Polish synsets lexically differentiated for gender.

\[
\text{(Example 11)}
\]

\[
\{\text{bondswoman 1, bondsman 2}\} \rightarrow \text{I-hyper} \rightarrow \{\text{gwarant 1, poręczyciel 1}; \{\text{poręczycielka 2}\}
\]

6.1.3 Other marked forms

Apart from lexically marked gender, Polish has a variety of other marked forms such as diminutives and augmentatives, which either do not appear or are very rare in English. plWordNet has a special relation of markedness (Polish nacechowanie) to show the links between base forms and their derivatives. Crucially, it is a relation between lexical units, not synsets. It has three variants: istota młoda ‘young creature’, diminutywność ‘diminutiveness’ and augmentatywność i ekspresywność ‘augmentativeness and expressiveness’. Lexical units in Polish which denote young creatures but are not derivative forms, such as ciele or cielak ‘calf’,

\[^5\] In the process of mapping of plWordNet to PWN we make an effort to repair dictionary-content gaps in plWordNet and make a list of such gaps in PWN (just in case they can be repaired in the future).
prosię or prosiak ‘piglet’, are linked to {młodzik, młodziak 2} ‘young animal’ via hyponymy. Analogically, in PWN synsets denoting young animals are attached by hyponymy to synsets denoting young sub-kinds of animals, such as {young mammal 1}. Now, PWN often places lexical units denoting young animals with diminutive forms, when such forms exist, as in, for instance, {kitten 1, kitty 3}, or {piglet 1, piggy 1, shoat 1, shote 1}. Since lexical units denoting young creatures and diminutive lexical units are not always in the same synsets, they are linked to their corresponding PWN synsets via I-hyponymy relation, as illustrated below:

(Example 12)  {prosiaczek 1} — dimin → {prosiak 1}
{kitten 1, kitty 3}, {piglet 1, piggy 1, shoat 1, shote 1} — hypo → {young mammal 1}
{piglet 1, ..} — I-hypernymy → {prosiak 1, prosię 2}; {prosiaczek 1, prosiątko 2}

In the rare cases without direct equivalents, I-synonymy will be applied as an obvious choice. For the remaining cases which lack English equivalents, we have decided to use I-hyponymy to link them to their English hypernyms.

6.2 Structural differences

6.2.1 Synonymy and synsets

The different strategy of synset construction and the resulting different idea of intra-lingual synonymy have led to systematic discrepancies in the structure of synsets in plWordNet and PWN. To begin with, plWordNet systematically distinguishes between count and mass nouns and never places them in the same synset. Conversely, PWN often neutralises the mass/count distinction at the synset level, putting both mass and count lexical units into one synset (furniture 1, piece of furniture 1, article of furniture 1) (cf. Miller 1998: 36). Such cases may cause problems for mapping, because it is hard to determine which plWordNet synset should be linked via I-synonymy, if any. Another option would be to resort to I-hyponymy to link the corresponding count- and mass noun plWordNet synsets:

(Example 13)  {mebel 1} — I-hypo → {furniture 1, piece of furniture 1, article of furniture 1}
{mebel 1} — hypo → {element wyposażenia 1}, {sprzęt 2}
{mebel 1} — meronymy → {umeblowanie 1}
{umeblowanie 1} — I-hypo → {furniture 1, piece of furniture 1, article of furniture 1}

There are also PWN synsets with singular and plural forms of the same lemma, such as {dumpling, dumplings 1} which have singular and plural hyponyms such as {matzo ball 1}, {wonton 1}, {gnocchi 1}. These are also linked via I-hyponymy to their corresponding plWordNet synsets:

(Example 14)  {pierog 1, pieróg 2} — I-hypo → {dumpling 1, dumplings 1}
{pierogi 1} — I-hypo → {dumpling 1, dumplings 1}
{matzo ball 1}, {wonton 1}, {gnocchi 1} — hypo → {dumpling 1, dumplings 1}

The differently defined synonymy affects the definition of hyponymy in plWordNet and PWN. In PWN, singular and collective nouns (pluralia tantum) may be hyponyms/hypernyms of each other. This is impossible in plWordNet: {dumpling 1, dumplings 1} ‘small balls or strips of
boiled or steamed dough’ is a hypernym of synsets {gnocchi} ‘Italian) a small dumpling made of potato or flour or semolina that is boiled or baked and is usually served with a sauce or with grated cheese’, {matzo ball, matzoh ball, matzah ball} ‘a Jewish dumpling made of matzo meal; usually served in soup’ and {won ton, wonton} ‘a Chinese dumpling filled with spiced minced pork; usually served in soup’.

A somewhat drastic, though maybe not unmotivated, case of using a broad notion of synonymy in PWN is the synset {monte 1, four-card monte 1, three-card monte 1} ‘a gambling card game of Spanish origin; 3 or 4 cards are dealt face up and players bet that one of them will be matched before the others as the cards are dealt from the pack one at a time’. It is obvious that a four-card monte is not a synonym of a three-card monte, they are just both hyponyms of monte. In Poland monte is not so popular. There only is a three-card monte – trzy karty (literally ‘three cards’). The synsets were joined by inter-language hyponymy, since the English equivalent three-card monte of the Polish lexical unit is within the set of the PWN synset:

(Example 15)

{trzy karty 1} —I-hypo→ {monte 1, four-card monte 1, three-card monte 1}

To sum up, we have decided to adopt a consistent strategy and use I-hyponymy in all cases of mixed PWN synsets.

6.2.2 Differently defined relations

Although there is a lot of correspondence between the set of linguistic relations employed by PWN and plWordNet and their respective construction, there are differences. They are reflected in the structure of both wordnets and may have consequences for the mapping. To give an example, PWN uses the conjunction or in its definitions, thus allowing for the hypernymy and/or, while plWordNet restricts its hypernymy to and. For example, the PWN synset {musical 1, …} was given the gloss ‘a play or film whose action and dialogue is interspersed with singing and dancing’; it received the following relational description in PWN (two instances of hyponymy):

{musical, musical comedy, musical theater} —hypo→ {movie, film, picture, …},

{musical, musical comedy, musical theater} —hypo→ {play 2}.

The word musical gained a similar definition in the Universal Dictionary of Polish (UDP, Dubisz 2004): ‘widowisko teatralne lub filmowe o treści komediowej lub melodramatycznej złożone z partii mówionych, śpiewanych i tanecznych’ [a theatre or film spectacle with comedic or melodramatic content, consisting of oral, sung or danced parts]. We had to split the concept into theatrical musical and musical film in order to avoid or-hyponymy:

{musical 1, komedia muzyczna} —hypo→ {film 1, obraz 6} ‘movie, picture’,

{musical 2} —hypo→ {przedstawienie 7} ‘play’

Or-hyponymy was banned from Polish WordNet in order to preserve the transitivity of hyponymy.6 Let us take a closer look at the English synset {musical 1, …}. It also contains

6 Transitivity seems to be fully accepted by general linguists feature of hyponymy: “We are not aware of any convincing examples of a genuine breakdown in the transitivity of hyponymy” (Croft and Cruse 2010: 145).
synonyms musical comedy and musical theatre. The first of those is a synonym of musical, according to Merriam-Webster Dictionary Online. The second one clearly refers to theatrical musical, according to Oxford English Dictionary. In fact the latter lexical unit should be a hyponym, not a synonym, of musical in the broader sense. This leads to a paradox: two synonyms of the synset have both hyponymy relations (to a play and to a film), while music(al) theatre has only one (to a play). The opposite could be noted in plWordNet where the lexical unit komedia muzyczna could be found in the meaning film musical. It is linked to {film 1, obraz 6} ‘movie, picture’ with hyponymy and is, of course, a synonym of Polish musical 1. It seems that in PWN hyponymy is only partly transitive and in some cases synonymy captures cases of hyponymy. Musicals from PWN and plWordNet had to be, naturally, linked with I-hyponymy:

(Example 16)    Pol. {musical 1, komedia muzyczna} —I-hypo→ Eng. {musical 1, …}

Pol. {musical 2} —I-hypo→ Eng. {musical 1, …}

Glosses
Not only relations cause difficulties. Glosses are also a source of various dilemmas for linguists who work on mapping. A case of thriller ‘a suspenseful adventure story or play or movie’ is somehow similar to musical. Here the connective or appears twice, surprisingly followed by only one hyponymy:

{thriller 1} —hypo→ {adventure story 1, heroic tale 1}

The UDP (Dubisz 2004) gives a very similar definition: ‘film, powieść lub sztuka teatralna, których sensacyjna i pełna napięcia akcja nasycona jest elementami tajemniczości, grozy i niesamowitości; dreszczowiec’ [a movie, novel or play whose sensational and supenseful action is imbued with elements of mystery, horror and eeriness; a thriller]. In plWordNet we distinguished three senses according to the medium: (a) ‘movie’, (b) ‘novel’, (c) ‘play’. Unfortunately the English gloss does not fit the English relation instance for thriller, so we are not sure whether Polish {thriller 1} should be linked to English {thriller} with I-synonymy (according to the English relation instance) or whether all three Polish thrillers ought to be connected with the English synset with I-hyponymy (according to the English gloss). We chose the former solution, assuming that relations are more important for a wordnet than glosses:

(Example 17)    Pol. {thriller 1} ‘novel’ —I-syn→ Eng. {thriller 1},

Pol. {thriller 2, dreszczowiec 1} ‘movie’ —I-hypo→ Eng. {movie, film, picture, …},

Pol. {thriller 3} ‘play’ —I-hypo→ Eng. {play 2}

6.2.3 Different relations to code the same conceptual dependencies
Sometimes equivalent synsets are interlinked within the two wordnets in a different way. For instance, the synset {jewelry, jewellery} is linked via hyponymy to synsets {bracelet, bangle} ‘jewelry worn around the wrist for decoration’, {ring, band} ‘consisting of a circlet of precious metal (often set with jewels) worn on the finger’, {earring} ‘jewelry to ornament the ear; usually clipped to the earlobe or fastened through a hole in the lobe’ and many others. The Polish counterpart of jewellery – biżuteria – is linked to the equivalents of ring (pierścionek, pierścień), bracelet (bransoletka) and earring (kolczyki, klipsy) and so on. Linking is with a different relation type, holonymy (the part subtype). So, the equivalent synsets are involved in different structures (hyponymic and meronymic):
6.2.4 Different lexico-semantic relations and different synset relations

We have already discussed (in section 6.1.3.) the plWordNet relation of markedness in plWordNet which do not have any exact counterpart in PWN (there is in fact a more general is-a-derivative-of relation, but it is too broad for our purposes). There are a few more nominal relations specific for plWordNet. The inhabitant relation is quite instructive.

In PWN the synset {American 1} ‘native or inhabitant of the United States’ is linked to the synset {United States, United States of America, America, the States, US, U.S., USA, U.S.A.} via the member meronymy relation. In our wordnet {Amerykanin 2} and {USA, Ameryka, Stany Zjednoczone, Stany} are connected via the inhabitant relation, because Amerykanie (Americans plural) inhabit Ameryka and this is expressed by the systematic derivational relation (Maziarz et al. 2011). Despite this difference, the synsets will be linked via I-synonymy:

(Example 18)

\{
{United States, ...} \leftrightarrow \text{I-synonymy} \rightarrow \{USA, Ameryka, Stany Zjednoczone, Stany\}
\}

\{American 1\} \leftrightarrow \text{I-synonymy} \rightarrow \{Amerykanin 2\}

6.2.5 Differently distinguished senses

The two wordnets differ not only in the repertory of lexico-semantic relations but also in sense distinctions. We have already discussed the case of English thriller which is strongly interrelated with three ‘thrillers’ in Polish WordNet, and the word musical which in plWordNet gained two equivalent musicals (its I-hyponyms). The case of chapel and kaplica is similar. In PWN chapel was given a definition ‘a place of worship that has its own altar’. Because in plWordNet senses are distinguished by relations in which a particular word-sense pair is involved\(^7\), Polish kaplica had two meanings: ‘autonomous building with its own altar’ and ‘part of another building (church or cloister) with its own altar’. Relation instances of the two synsets are illustrated below:

\{kaplica 1\} ‘chapel, autonomous building’ — hypo → \{świątynia\} ‘temple’

\{kaplica 1\} — hyper → \{kaplica przycmentarna, ...\} ‘cemetery chapel’

\{kaplica 2\} ‘chapel, part of another building’ — hypo → \{pomieszczenie 3\} ‘room’

\{kaplica 2\} — meronymy:place → \{klasztor 1\} ‘monastery’

\{kaplica 2\} — meronomy:place → \{kościół 2\} ‘church’

The two senses do have different lexical neighbourhoods, so we assume that they should stay separate. PWN shows an alternative way of describing the concept ‘chapel’ – instead of splitting the sense it was kept intact and linked to a higher hypernym \{place of worship, ...\}. At a first glance the two approaches appear justified.

\(^7\) According to (Croft and Cruse 2012: p. 113) one of sense difference symptoms are distinct sets of lexico-semantic relations for considered lexical units.
Unfortunately, the hypernym \{place of worship, house of prayer, house of God, house of worship\} was itself linked to \{building, edifice\} and was given too narrow a definition ‘any building where congregations gather for prayer’, although \{chapel\} has two hyponyms that clearly are not buildings: \{lady church\} ‘a small chapel in a church; dedicated to the Virgin Mary’ and \{side chapel\} ‘a small chapel off the side aisle of a church’. Despite this inconsistency we decided to link our \{kaplica 1\} and \{kaplica 2\} with \textit{I-hyponymy} with \{chapel 1\}, assuming that it has both meanings:

\begin{enumerate}
  \item \{kaplica 1\} ‘building’ — I-hypo \{chapel 1\} ‘a place of worship’,
  \item \{kaplica 2\} ‘room’ — I-hypo \{chapel 1\} ‘a place of worship’.
\end{enumerate}

6.2.6 Dictionary-content mismatches

Mapping is also made more difficult by \textit{dictionary content gaps}. We have decided that, though we could improve \textit{plWordNet}, we were not supposed to make any changes inside PWN during the mapping process. What are dictionary gaps? Lexical gaps are caused by specificities of the two languages, dictionary gaps are produced by limitations of any dictionary/thesaurus/wordnet size. For instance, in PWN names of artists are restricted to only one domain of art even in cases when they apply to more than one art domain (quite systematically). For example, \{impressionist 1\} ‘a painter who follows the theories of Impressionism’ has one hypernym relation instance to \{painter 1\}, although there is a clear evidence that the word could be used also to indicate impressionist musicians – see entry in (Procter 1978) or poets – see \textit{impressionism in} (Myers, Wukasch 2003). Polish \textit{impresjonista} ‘impressionist painter, musician or poet’ is defined using two \textit{and}-hyponyms \textit{artysta} ‘artist’ and \textit{przedstawiciel} ‘exponent (of an artistic trend)’. We cope with the lexical database mismatch between PWN and \textit{plWordNet} simply using \textit{I-hyponymy} between more specific English \{impressionist 1\} and broader Polish \{impresjonista 1\}.

Conclusion and perspectives

The system of inter-lingual relations and the mapping procedure proposed in this paper have been shown to work successfully. We have managed to map about 18000 \textit{plWordNet} synsets onto PWN synsets. All edited \textit{plWordNet} synsets have been linked to PWN’s synsets by one of the proposed inter-lingual relations. The manual mapping was enhanced by the automatic prompt system, which turned out useful (according to linguists involved in the mapping). The created mapping is especially valuable in the light of the fact that we have been linking two completely independently created large-scale wordnets. It enabled a systematic comparison of \textit{plWordNet} and PWN’s structure and content, but also \textit{plWordNet}’s verification and correction. We have encountered mapping dilemmas boiling down to lexico-grammatical differences between English and Polish (e.g., lexical gaps, rich morphology of Polish) and to structural incompatibilities resulting from different methodologies lying behind the construction of the two wordnets (e.g., differently defined relations) to which we have proposed systematic solutions.
References


Polish Wikipedia: pl.wikipedia.org

Polish Wiktionary: pl.wiktionary.org


