COGNITIVE STUDIES | ÉTUDES COGNITIVES, 15: 335–351
Warsaw 2015

DOI: 10.11649/cs.2015.023

EWA RUDNICKA[A], **WOJCIECH WITKOWSKI**[B], & **MICHAŁ KALIŃSKI**[C]

G4.19 Research Group, Wrocław University of Technology, Poland
[A]ewa.rudnicka@pwr.edu.pl ; [B]wojciech.witkowski@uwr.edu.pl ;
[C]michal.kalinski@pwr.edu.pl

TOWARDS THE METHODOLOGY FOR EXTENDING PRINCETON WORDNET

Abstract

The paper presents the methodology and results of the first, pilot stage of the extension of Princeton WordNet, a huge electronic English language thesaurus and lexico-semantic network based on synsets, ie. sets of synonymous lexical units, or lemma sense pairs. The necessity for such extension arose in the course of mapping plWordNet (Polish WordNet — Słowosieć) onto Princeton WordNet, which produced a large number of inter-lingual hyponymy links signalling differences in the structure and lexical coverage of the two networks. The proposed strategy uses I-hyponymy links as pointers to presumed gaps in the lexical coverage of Princeton-WordNet and offers strategies of filling them in with new lexical units and synsets.
Keywords: wordnet extension; lexico-semantic relations; I-hyponymy links; rule-based algorithms; lexical gaps

1. Introduction

Of the many available electronic lexical resources, Princeton WordNet (henceforth, PWN) is probably one of the most exceptional ones. This is because it is not yet another electronic dictionary of a standard format. It is much more. The resource combines the type of data genuine to a corpus-based monolingual dictionary as well as an extended thesaurus, and, above all, a lexico-semantic network organised in a machine-readable format (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Due to the latter, it is widely used by computational linguists and natural language engineers for the purposes of various application tasks in the areas of natural language processing (henceforth, NLP), natural language engineering and technology (e.g. Word Sense Disambiguation, Information Retrieval).[1]

[1]The list of major applications of Princeton WordNet can be found on the official Princeton WordNet website: http://wordnet.princeton.edu/wordnet/related-projects/

Essential as its unique format is, Princeton WordNet would not be applicable in NLP tasks, if it were not for its size, which puts it among the biggest existing resources for the English language. The latest 3.1 version contains 155 593 word forms (*lemmas*), 206 978 word meaning pairs (*lexical units*), and 117 659 synonym sets (*synsets*). They are linked by a rich set of lexico-semantic relations such as hyponymy/hypernymy, meronymy/holonymy, antonymy etc. Their instance counts amount to 207 269 for lexical units relations and for 117 790 synset relations. This is undoubtedly a remarkable size, although, as in the case of every lexical resource, there is always room for improvement, especially from the perspective of large scale processing tasks.

For such type of resource to serve its functions, it should be constantly monitored and developed. Unfortunately, it is no longer the case of Princeton WordNet. Due to the conclusion of research projects and the subsequent break-up of the team, works on the further construction of PWN have been recently suspended. The latest 3.1 version made it to the public in 2012, but its lexical content was not much enriched in comparison to the previous 3.0 version, whose release dates back already to 2006.[2] It is actually 3.0 version which is still more commonly used worldwide, both for the purposes of mapping to other languages' wordnets and for different application tasks[3] (e.g. EuroWordNet, MultiWordNet, IndoWordNet, FrameNet, SentiWordNet).

When in 2012 Polish WordNet (henceforth, plWN) team started the process of mapping plWordNet onto PWN, also its 3.0 version was used, but it was later transformed to 3.1.[4] Throughout all stages of mapping, the most frequently introduced relation between plWordNet and Princeton WordNet synsets has been inter-lingual hyponymy, introduced in cases of the absence of direct English equivalents (Rudnicka, Maziarz, Piasecki, & Szpakowicz, 2012; Maziarz, Piasecki, Rudnicka, & Szpakowicz, 2013). In the latest 2.2 version of plWordNet, its count doubles that of inter-lingual synonymy, both for noun and adjective synsets (Piasecki, Maziarz, Szpakowicz, & Rudnicka, 2014). Obviously, such high number of I-hyponymy links cannot be attributed to dictionary coverage gaps alone. A lot of those links are simply the result of the existence of classic 'lexical gaps' (Svensen, 2009). Still, the number of the observed dictionary coverage gaps is significant.

Both from the perspective of standard dictionary uses (such as e.g. translation and writing) as well as from the perspective of natural language processing tasks, I-hyponymy links are much less precise, and hence less desired, than I-synonymy links. With no chances for a new release of PWN in any foreseeable future, an idea of a limited extension of Princeton WordNet came up. The major aim of such an extension would be to fill dictionary gaps noted in the process of mapping plWordNet to PWN and substitute the existing I-hyponymy links with much more informative and useful I-synonymy links. The proposal which is going to be put forward in this paper relies on the very I-hyponymy links, which are used as pointers

[2] Christiane Fellbaum, p.c.

[3] see ft. 1 above.

[4] The analysis of 3.0 to 3.1 mapping has shown some minor changes between the two versions, mainly in spelling, synset content (lexical units deletion, addition, transposition) and their respective numbering. Occasionally, some new lexical units and synsets were introduced.

to the presumed Princeton WordNet gaps (missing lexical units) and even whole 'empty nests' (several missing co-hyponyms of one hypernym synset) in the network. The process is supported by an automatic translation of Polish synsets' lemmas by a large cascade dictionary and the filtering of the obtained translations by Princeton WordNet lemmas. The introduction of new lexical units and synsets is preceded by careful dictionary and corpora studies.

The paper is organised as follows: after the introduction offered in Section 1, we will provide background and motivation for the extension of Princeton WordNet in Section 2. The general extension strategy will be described in Section 3, while the results of the first, pilot stage of the extension process will be discussed in Section 4. The paper will close with conclusions in Section 5.

2. Princeton WordNet and its mapping to plWordNet

Princeton WordNet is the first resource of a wordnet type ever created, therefore it is sometimes simply referred to as WordNet. Its development process started at Princeton University in the mid 80-ties and continued up to 2006, with some minor alterations still made by 2012. WordNet originated as an experiment on mapping lexical memory of children, yet it gradually evolved into a huge electronic resource mapping the large part of the lexical system of English (Fellbaum, 1998). Unfortunately, works on the further development of WordNet have recently stopped (see Section 1 above). WordNet quickly found its followers such as, for instance, GermaNet for German (Hamp & Feldweg, 1997). Subsequently, works followed on the connection of all those newly built resources to the original Princeton Wordnet. The outcomes are such multi-lingual wordnets as EuroWordNet (Vossen, 2002), MultiWordNet (Pianta, Bentivogli, & Girardi, 2002), or IndoWordNet (Pushpak, 2010). Unfortunately, most of those wordnets were to a large extent constructed by an automatic 'transfer and merge' method consisting in the translation of PWN content and structure to other languages.

plWordNet is one of the very few world wordnets built wholly manually and largely independently of Princeton WordNet by a team of linguists and lexicographers supported by language technology tools extracting data from a large corpus (Piasecki, Szpakowicz, & Broda, 2009; Maziarz, Piasecki, & Szpakowicz, 2012). It is currently the biggest existing wordnet. With 159 091 lemmas, 225 758 lexical units, and 168 663 synsets it overgrows even PWN (counts as of 23th January 2015).

The mapping of plWordNet to Princeton WordNet started in 2012 with the category of nouns (Rudnicka et al., 2012). At this point plWN already reached a pretty mature shape, so the risk of influencing the structure of plWN by that of PWN was small. A detailed mapping procedure and a set of seven hierarchically ordered inter-lingual relations were defined (see Appendix). The mapping procedure consists of three major steps: recognising the sense of a source synset, finding a target synset and linking the source synset with the target synset by means of the most appropriate inter-lingual relation. The set of inter-lingual relations for noun mapping comprises synonymy, partial synonymy, inter-register synonymy, hyponymy, hypernymy, meronymy and holonymy. The mapping direction is from plWN to PWN, for which a bottom-up approach (going from the lowest levels of

the hypernymy hierarchy up to the higher ones) has been adopted. The manual mapping process is supported by a specially designed automatic prompt system (Kędzia, Piasecki, & Przybycień, 2013). Recently, the works on the mapping of adjectives have commenced.

From the very beginning of works on mapping plWordNet onto Princeton WordNet, a number of substantial differences in the lexical coverage of the two resources was noted (Rudnicka et al., 2012). They were the primary reason for the introduction of seven inter-lingual relations. Still, already the first mapping results demonstrated the 'supremacy' of certain relations over others in terms of frequency. Interestingly, it was inter-lingual hyponymy relation that has always ruled out over inter-lingual synonymy, although it is the latter which is the topmost one with respect to the priority of introduction. In an attempt to explain such state of affairs, an analysis of the mapping results was conducted. The main reasons turned out to be the existence of classic lexical gaps between English and Polish and differences in the dictionary coverage as well as in the structure of relations between plWN and PWN. The tendency for the dominance of I-hyponymy over I-synonymy has prevailed up to now and the current counts of the inter-lingual relations are presented in Table 1 below (data as of 23th January 2015).

Table 1: Inter-lingual relations counts in plWN

Relation	**Nouns**	**Adjectives**
I-synonymy	30476	3351
I-partial synonymy	2896	1113
I-inter-register synonymy	1564	38
I-hyponymy	61765	10280
I-hypernymy	3919	46
I-meronymy	6200	0
I-holonymy	1659	0
I-cross-categorial synonymy	0	6900

The most striking piece of data from Table 1 is definitely the number of I-hyponymy links which in case of nouns doubles the number of I-synonymy links, while in case of adjectives it overgrows it three times. One easy explanation is the existence of a presumably high number of lexical gaps between English and Polish. Nevertheless, it cannot account for so many cases. It is a very undesired result from the perspective of applications of a bilingual lexical resource (such as, for instance, bilingual word sense disambiguation), since I-hyponymy links are much less precise and informative than I-synonymy links. Therefore, we decided to make an attempt at verifying the origin of such vast number of I-hyponymy links. They were assumed to be pointers to the presumed gaps in the lexical coverage of PWN. A strategy for identifying genuine gaps in the lexical coverage of Princeton WordNet and filling them in with new lexical units and synsets was developed and will be presented in Section 3 below.

3. General extension strategy

Since the main motivation for the extension of Princeton WordNet has been the high percentage of inter-lingual hyponymy links holding between plWordNet and PWN synsets, it was a natural move to make use of the very links in designing an extension strategy. They are taken as the input for the first stage of strategy which aims at detecting lexical coverage gaps in Princeton WordNet. The set of I-hyponymy links to be used is limited to those holding between plWN synsets located at the lowest levels of Polish hypernymy hierarchy. It is predicted that there is a good chance that their direct equivalents (if such exist) will also be located at the lowest levels of I-hypernymy hierarchy in PWN. For this first, pilot extension of PWN, we have decided to introduce new PWN synsets only at the lowest levels of PWN hypernymy hierarchy in order not to modify the structure of the original Princeton WordNet. Last but not least, the extension process will be limited to nouns. Nouns are the standard starting point in any wordnet construction, as the 'easiest' and the most numerous category. Moreover, the relation structures of nouns in plWN and PWN are parallel to a large extent.

Now, the general scheme of the proposed extension strategy is as follows. Lemmas of the selected plWN synsets are automatically 'translated' by a large cascade dictionary.[5] Next, the obtained 'translations' are filtered by the list of PWN lemmas. The results are divided into three groups. The first group encompasses lemmas of Polish synsets for which the cascade dictionary found translations/equivalents whose lemmas are absent from PWN. The second group consists of Polish lemmas for which the dictionary did not find any matching translations. Finally, the third group includes lemmas for which the dictionary found translations whose lemmas are present in PWN. Those lists are treated as the basis for the extension of PWN.

In the second stage, the actual extension of PWN takes place. Lexicographers start their work with the first group which contains suggested English equivalents whose lemmas are absent from PWN. Every suggestion for a new lexical unit is carefully verified. First, lexicographers refer to their linguistic knowledge and consult dictionaries in order to confirm that there exists a sense of the suggested lemma that directly matches that of a plWN synset from the list. Next, they check the frequency of occurrence of the given lemma in the frequency and entropy lists generated from the XML edition of *British National Corpus* (BNC). The minimal frequency required for a lemma to be introduced into PWN is 5 hits. When the lemma is absent from BNC, lexicographers resort to Google search of British and American web pages. In such cases the minimal required frequency is set to 10 hits.

The location of new synsets in the wordnet graph structure is largely determined by inter-lingual hyponymy links which lead to a specific English hypernym 'nest'. New English synsets are added as hyponyms of PWN synsets functioning as inter-lingual hypernyms. The last step for such addition is reading through an intra-lingual hyponymy relation test adopted from EuroWordNet (Vossen, 2002) and cited below:

[5]The cascade dictionary combines several traditional dictionaries, among them (Piotrowski & Saloni, 2002) and a large proprietary dictionary of TiP company. For the purposes of translation, their data were ordered in the hierarchy of importance; the topmost gaining more priority.

Test 1		**Hyponymy-relation between nouns**
yes	a	A/an X is a/an Y (with certain properties) It is a X and therefore also a Y If it is a X then it must be a Y
no	b	the converse of any of the (a) sentences.
Conditions:		— both X and Y are singular nouns or plural nouns

The use of EuroWordNet intra-lingual tests was motivated by the fact that such tests do not appear in the literature on PWN construction (Miller et al., 1990; Fellbaum, 1998). Moreover, since EuroWordNet was designed to integrate European languages of various lexico-semantic properties, the tests for intra-lingual relations exhibit a noticeable degree of language independence. Lastly, each new PWN synset is followed by a definition and usage examples, as in the original Princeton WordNet. One of the main sources for definitions is English Wikipedia. The examples are taken from reliable, open-access English sources.

Once lexicographers work through the first group of plWN lemmas and their suggested English equivalents, they proceed to the second group which encompasses plWN lemmas for which the cascade dictionary did not find any equivalents. Obviously, it does not mean that they are necessarily missing in the English language. Their absence may be due to the content limitation of the lexical resources used. In any case, lexicographers have to search for equivalents themselves consulting available lexical resources such as various types of dictionaries (including Wikipedia), corpora and reliable Internet sources. They are informed that a non-trivial number of Polish lexical units may not have their equivalents in English, their lack being persistent to the existence of classic lexical gaps (Svensen, 2009). Lexicographers are also warned not to generate 'artificial equivalents', but to introduce only genuine English vocabulary, therefore every equivalent proposal has to undergo the same verification path as in the case of the suggested translations from the first group.

The last, third group is the most challenging one. It includes suggested dictionary equivalents whose lemmas are already present in PWN, but their respective synsets are not linked via inter-lingual synonymy relation to plWN synsets in question (those whose lemmas are on the list). In this case the lexicographer has to check the validity of the existing inter-lingual relation and, if they deemed it justified, introduce changes. Sometimes the network will stay as it is, sometimes it will undergo alterations. In some cases new PWN synsets will also be introduced following the procedures for the first and second group. All changes are recorded in a special document. They include changing the existing inter-lingual relation, adding a new lexical unit (and synset) with a new sense of the already existing lemma, adding a completely new lemma. Cases when no changes are introduced to the network of relations usually signal the existence of lexical gaps on the English side.

4. Results

4.1. Results for the selected semantic domains

Following the proposed extension strategy, the first step in the extension process is

filtering out plWN synsets that are, first, linked via inter-lingual hyponymy relation to PWN synsets, second, located at the lowest level of the hypernymy strategy. Table 2 presents the counts of the very plWN synsets ordered by their number and semantic domain.

Table 2: The counts of plWN 'leaf' synsets linked by I-hyponymy relation to PWN synsets.

Synset qualifier	no. of plWN synsets	Synset qualifier	no. of plWN synsets
[wytw] [man-made artifacts]	7657	**[jedz] [food]**	**885**
[zw] [animals]	6616	[zwz] [human activity related ideas]	799
[os] [person]	6361	[sys] [systematic names]	763
[msc] [places]	2789	[pos] [possession]	658
[rsl] [plants]	2736	**[il] [quantities]**	**547**
[umy] [mental activity]	1854	[czc] [body parts]	395
[por] [communication]	1774	[st] [states]	316
[cech] [properties]	1492	[zdarz] [events]	306
[prc] [processes]	1382	[czas] [time]	262
[rz] [natural objects]	1377	**[ksz] [shapes]**	**156**
[zj] [(natural) phenomena]	1242	[czuj] [feelings]	155
[sbst] [substances]	1181	[cel] [aim of action]	18
[grp] [groups]	1064		

For the first, pilot stage of the extension process we have selected three semantic domains: food, quantity and shape. The number of plWN leaf synsets (holding I-hyponymy relation) from all these domains amounts to 1558.

Figures in Graphs 1 and 2, which are presented below, illustrate the relative proportions of the domains for the selected and the introduced synsets. As can be seen in these graphs, when the proportions of the synsets that were selected are compared to the proportions of the synsets that were actually added to PWN, no noticeable differences differences among the three groups can be observed.

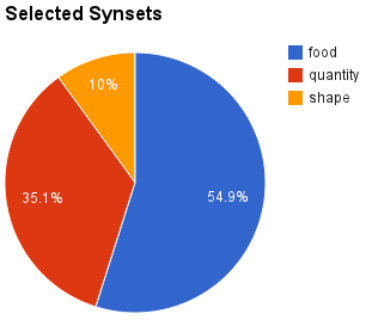

Figure 1: Proportions of domains in the set of selected synsets

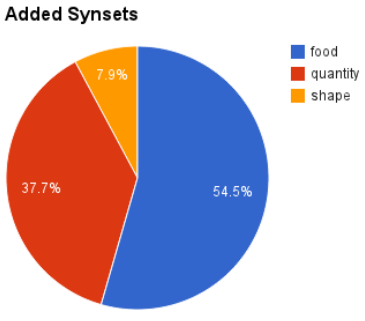

Figure 2: Proportions of domains in the set of introduced synsets

The total number of newly introduced synsets equals 751. This equals to 48% of the total number of synsets in the selected domains. The figures of introduced synsets for the relevant domains are as follows: food — 409 out of 855 (47%), quantity — 283 out of 547 (52%) and shape — 59 out of 156 (38%). These figures are presented graphically in Graph 3.

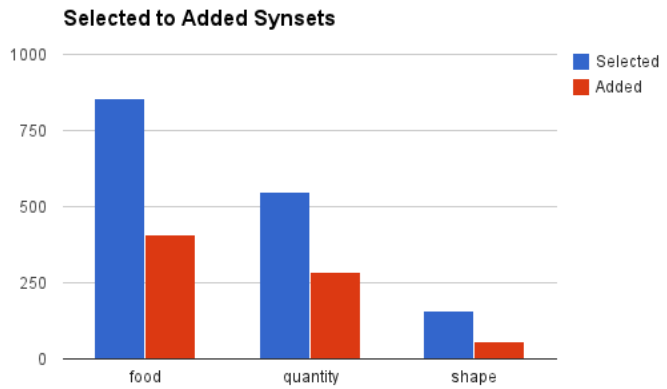

Figure 3: Comparison of figures for selected and introduced synsets.

The results of the comparison of proportions of selected and added synsets coupled with the results of the comparison between the numbers of selected synsets and the synsets that were actually added to PWN may be treated as an insight into the possible shape of PWN if the extension procedure were to be carried in all domains. Accordingly, it could be expected that PWN database could be expanded by roughly 50% of the overall number of the potential synsets selected from plWN. Moreover, the expansion of relevant domains within these 50% of new synsets would exhibit size proportions that are present in plWN resources.

4.2. New synsets introduced

Examples of synsets added to PWN resources are presented below in Tables 3, 4, 5. These examples are divided in accordance to the division of synsets into three groups based on the presence of English equivalents in the cascade dictionary and PWN resources.

The examples presented in Tables 3, 4, 5 were added into PWN resources on the basis of the extension procedures that were devised to account for each of the selected groups of synsets respectively. As far as the detailed steps in these procedures are concerned, the reader is referred to Appendix.

Table 3: Examples of introduced Group 1 synsets.

plWN synset[6]	**Suggested lamma**	**Number of occurrences**	**Added synset**
[centylion 1 (il)] (liczba 100^{600} czyli jedynka i sześćset zer *number* 100^{600} *that is one and six hundred zeros*)	centillion	over 10000 (Internet)	[centillion 1 (il)] (a number, which is equal to either 1 followed by 303 zeros, or 1 followed by 600 zeros, depending on the system used)
[tuszonka 1 (jedz)] (konserwa mięsna, mielonka wieprzowa *canned meat, stewed pork*)	tushonka	over 99000 (Internet)	[tushonka 1 (jedz)] (Tushonka is a kind of canned stewed meat especially popular in the CIS and other countries of the former Soviet Union)
[izoklina 1 (ksz)]	isocline	over 41000 (Internet)	[isocline 1 (msc)] (a curve through points at which the parent function's slope will always be the same, regardless of initial conditions)

Table 4: Examples of introduced Group 2 synsets.

plWN synset	**Lemma chosen by lexicographer**	**Number of occurrences**	**Added synset**
[pensum 1(il)] (określona liczba obowiązkowych do przepracowania godzin przez pracownika oświaty) *a specified amount of obligatory working hours for a teacher*	teaching quota	419 (Internet)	[teaching quota1(il)] (the minimum amount of classes a teacher has to conduct, as stated in his contract)
[trawa cytrynowa 2 (jedz)] orientalna przyprawa, dodatek do dań, listki, pędy (czasem przetworzone) rośliny nazywanej tak samo *oriental spice made from the plant of the same name which is added to various dishes*	lemon grass	around 2320 (Internet)	[lemon grass 3 (jedz)] (Cymbopogon citratus, commonly known as lemon grass or oil grass, is a tropical plant from Southeast Asia)
[bróg 1 (ksz)]	hayrack	around 107000 (Internet)	[hayrack 1 (wytw)] (a rack that holds hay for feeding livestock)

[6]In the examples below, when no description is provided for the plWN synset, it means that no such description is available in Słowosieć database.

Table 5: Examples of introduced Group 3 synsets.

plWN synset	added PWN synset	Comment
[stopa depozytowa 1 (il)] (określa oprocentowanie jednodniowych depozytów składanych przez banki komercyjne w banku centralnym *it specifies the interest rate of one-day deposits placed by commercial banks in a central bank*)	[bank rate 1 (il)] (the discount rate fixed by a central bank)	it was possible to replace the established inter-lingual hyponymy relations with inter-lingual synonym relations after the addition of relevant synsets to PWN
[dzwonko 1 (jedz)]	[fish steak 1 (jedz)] (cross-section slice of a large fish)	
[profil 2 (ksz)] (linia obwodząca zarys kształtu czegoś, kontur *a line depicting the shape or a contour of sth*)	[profile 2 (ksz)] (an outline of something (especially a human face as seen from one side))	

4.3. No new synsets introduced

The first and most obvious explanation for the fact that about a half of the inter-hyponymy links did not lead to the introduction of new plWN synsets is the existence of lexical gaps between English and Polish. Such cases are noted for all the selected semantic domains (food, quantity, shape) and given in Table 6 below:

Table 6: Examples of Group 1 synsets that were not introduced.

plWN synset	Current I-ling relation	target PWN synset	Comment
[kapka 2 (il)] (odrobina, bardzo mała ilość czegoś (zwykle płynnego)) *a very small amount of sth usually liquid*	Inter-lingual hyponymy	[drop 2 (il)] (a small indefinite quantity (especially of a liquid))	no better English equivalent could be found
[bryzol 1 (jedz)] (delikatny płat mięsa rozbity i usmażony *a tender crushed and fried chop of meat*)	Inter-lingual hyponymy	[beefsteak 1 (jedz)] (a beef steak usually cooked by broiling)	no better English equivalent could be found
[izohigra 1 (ksz)] (linia łącząca punkty o jednakowej wilgotności *a line connecting points of identical humidity*)	Inter-lingual hyponymy	[isogram 1 (msc)] (a line drawn on a map connecting points having the same numerical value of some variable)	no better English equivalent could be found

Table 7: Examples of Group 2 synsets that were not introduced.

plWN synset	Suggested lemma	Number of occurrences	Comment
[łut 2 (il)] odrobina; tylko w wyrażeniu „łut szczęścia" *a bit, found only in an expression 'a bit of luck'*	lot unit		the suggested English lemma is invalid, Polish lemma [łut] is the part of a fixed expression
kopytka 1 (jedz) (potrawa mączna z dodatkiem ziemniaków i jaj *a flour, potato, egg based dish*)	kopytka	12700 (Internet)	synset was not added as the sources were Polish blogs written in English and hosted on UK servers

However, there are also other reasons for not introducing a new PWN synset on the basis of I-hyponymy link. As is illustrated in Table 7 above, new synsets were also not introduced, if the lexicographers encountered cases in which Polish lemma were frozen in fixed expressions or were found only in sources that were not created by native users of English language.

Of the three selected semantic domains, the total number of plWN synsets from Group 3 was 326. Of those 326 synsets, 40 were judged by lexicographers as holding valid interlingual hyponymy relations to PWN synsets. Example of such cases are illustrated in Table 8 below:

Table 8: Examples of Group 3 synsets that were not introduced.

plWN synset	PWN synset	Comment
[produkt krajowy brutto 1 (il)] (pojęcie ekonomiczne oznaczające jeden z podstawowych mierników dochodu narodowego stosowanych w rachunkach narodowych *an economic term defining one of the basic measurements of domestic product, which is used in national accounting*)	[gross domestic product 1 (il)] (total market values of goods and services produced by workers and capital within a nation's borders during a given period (usually 1 year))	established interlingual hyponymy relations are valid, no new synsets need to be added
[marynata 3 (jedz)] zalewa na bazie octu do marynowania przetworów, by zachowały trwałość *a vinegar based mixture used to conserve food*	[marinade 1 (jedz)] (mixtures of vinegar or wine and oil with various spices and seasonings; used for soaking foods before cooking)	
[bruzda 3 (ksz)] rowek pozostający po wyoraniu i odłożeniu skiby podczas orki *a trench left after ground has been ploughed*)	furrow 1 (wytw) a long shallow trench in the ground (especially one made by a plow)	

Clearly, the meaning of the plWN synset {marynata 3} is more specific than that of PWN synset {marinade 1}. A direct meaning equivalent of {marynata 3} does not exist in English, hence the I-hyponymy relation is the best one to be introduced. Such cases are examples of classic lexical gaps between English and Polish.

4.4. Problems

The most apparent problem was the number of lemmas denoting concepts related strictly to Polish domain. These synsets were especially problematic when they were related to unit, which was present in the PWN database, and that could etymologically be traced back to language other than Polish. An example of such a case are plWN synsets [korzec], [szefel] and PWN synset [bushel]. [korzec] is defined as an old Polish dry measure, similar to [bushel]; [szefel] is etymologically related to German *scheffel*, whose English equivalent is *bushel*. Accordingly, both [korzec] and [szefel] could be mapped onto PWN [bushel]. However, the frequency search for [korzec] and [szefel] has shown that, contrary to German [scheffel], these units do not occur in sources that can be treated as not strictly belonging to Polish language domain.

Another problem encountered by the lexicographers was the choice of the appropriate target lemma, in the cases where more than one potential English equivalent could match the Polish lemma. A case at hand may be the plWN synset [powietrznia 2] meaning 'the capacity of elastic arteries'. In this case, the Internet search left the lexicographer with the choice between *arterial distensibility* ('a measure of the arterial ability to expand and contract with cardiac pulsation and relaxation') or *compliance* ('the ability of a hollow organ (vessel) to distend and increase volume with increasing transmural pressure or the tendency of a hollow organ to resist recoil toward its original dimensions on application of a distending or compressing force'), which were both judged as the potential inter-lingual synonyms for the Polish synset [powietrznia 2]. The problem with the rejection of one of the lemma resulted from the lack of specialist knowledge in the field of medicine. Eventually, the synset [compliance 4] was added to PWN as the inter-lingual synonym of [powietrznia 2].

Next, the introduction of synsets in Group 3 has revealed a number of interesting pitfalls. Out of 326 noted cases, 242 cases are situations in which the lexicographers decided that the existing interlingual Polish-English relation established between plWN and PWN synsets is not a valid one. In these 242 cases the invalidity of the existing inter-lingual hyponymy relation was due to the fact that that the lexicographers were able to provide a more direct English equivalent for the plWN source synset. The examination of these cases has shown that the two causes of the prevailing number of errors marked as requiring an establishment of an inter-lingual relation to a different PWN synset.

One of the reasons for establishing invalid hyponymy relations between plWN and PWN synsets that were already present in the databases were the results of dictionary searches combined with the available databases resources. Due to the lack of more direct PWN equivalents that could serve as inter-lingual synonyms for plWN synsets the lexicographers decided to search for the closest hyperonyms.

At the PWN expansion stage, the lexicographer conducted a wider dictionary / Internet search, which resulted in the introduction of a new synset into PWN. The introduced synset served as the inter-lingual synonym for the source plWN synset. Instances of such a change are presented below in Table 9.[7]

Table 9: Addition of new I-synonym synset into PWN.

plWN Synset	PWN synset	Current_Relation	Error type	Intended_Target
masa molowa 1 [masa jednego mola materii (mass of one mol of substance)]	**gram molecule 1** [the molecular weight of a substance expressed in grams] **metric weight unit 1** [a decimal unit of weight based on the gram]	syn plWN-PWN/hipo plWN-PWN & hipo plWN-PWN respectively	[diff.syns]	**molar mass 1** [the mass of a given substance (chemical element or chemical compound) divided by its amount of substance.
waga brutto 1	weight 1 [the vertical force exerted by a mass as a result of gravity]	hipo plWN-PWN	[diff.syns]	**gross weight 1** [the total weight of a product and its packaging]

The other reason why the established hyponymy relation between plWN and PWN synsets was incorrect is the fact that PWN database contains a synset which is a proper inter-lingual synonym for the plWN synset, which possesses an inter-lingual hyponymy relation. With respect to such cases, lexicographer chose to map the plWN synset to its inter-lingual synonym, which resulted in the change of the existing inter-lingual relations. An example illustrating this is presented below in Table 10.

Table 10: Change of inter-lingual mapping due to existence of PWN I-synonym synsets.

plWN Synset	PWN synset	Current_Relation	Error type	Intended_Target
parytet 1 [zasada równości proporcji dwóch lub więcej wielkości, określana prawnie a priori (an a priori legally regulated rule of equality of proportions)]	**proportion 1** [the quotient obtained when the magnitude of a part is divided by the magnitude of the whole]	hipo plWN-PWN	[diff.syns]	**parity 5** [functional equality]

[7] In Tables 9, 10, 11, the label [Current_Relation] means the relation that was established between plWN and PWN synsets before the works on expanding PWN resources were carried out.

The cases presented above point rather to the technical issues than to empirical aspects of WordNet design. The presence of cases such as those in Table 9 indicates that mapping choices are substantially dependent on the resources available in the lexical databases that are to be correlated. The amount of the available resources (or in the case at hand the limited size of the resources in PWN) forced the lexicographers to establish inter-lingual relations that did not fully represent the degree of equivalence between the mapped units, but were the best that could be established provided the relevant resources. The cases illustrated by examples in Table 10 may be considered as resulting from the individual differences in linguistic knowledge and experience of the lexicographers.

Lastly, 6 cases of mismatch in the relation networks between plWN and PWN were observed. The mismatches in the networks of relation may be seen as illustrating differences in the organization of the mental lexicons of speakers of Polish and English that have influenced the choices of lexicographers, who have originally established a hyponymy relation between the respectful plWN and PWN synsets. An example of such a case is presented in Table 11, where the original relation was the inter-lingual hyponymy between plWN synset [lemoniada 2] and PWN synset [containerful 1]. [lemoniada 2] is dominated by synset [napój 2] — (porcja napoju, może to być ilość, która mieści się w naczyniu lub ilość oferowana w sklepach, zapakowana do butelek lub kartonów (a serving of a drink, an amount that can be contained in a container, or is offered at stores in bottles or boxes), whereas [containerful 1] is dominated by [indefinite quantity 1] — (an estimated quantity). Neither [napój 2] nor [containerful 1] are related to each other by means of inter-lingual relations. What is more, [containerful 1] dominates the synset [bottle 2], which provides a more restricted meaning than [containerful 1]. As a result, based on the mismatches in the relation networks and the definitions of the synsets the lexicographer decided to map [lemoniada 2] onto [bottle 2] by means of inter-lingual hyponymy.

Table 11: Change of inter-lingual mapping due to relation network mismatch.

plWN Synset	PWN synset	Current_Relation	Error type	Intended_Target
lemoniada 2 [porcja lemoniady, puszka lub butelka z napojem, ale także np. szklanka, dzbanek (a serving of lemonade served in a can, bottle, etc.)]	**containerful 1** [the quantity that a container will hold]	hipo plWN-PWN	[diff.syns] & [network. mismatch]	**bottle 2** [the quantity contained in a bottle]

Assuming the above approach, a careful scrutinity of the hierarchical organization of plWN and PWN resources may be treated as a source of information

relevant to more theoretically oriented linguistic studies in the fields of semantics and psycholinguistics.

5. Conclusion

Of the three semantic domains selected for the pilot stage of the extension process, about 1500 inter-lingual hyponymy links resulted in the introduction of about 750 new PWN synsets. Such results clearly justify the need for the extension of Princeton WordNet. An analysis of cases in which no new PWN was introduced has been demonstrated to be due to the existence of a large number of lexical gaps between English and Polish.

It was shown that from the perspective of Polish and Princeton WordNets' design the Polish-to-English direction may result in a beneficial procedure that allows to reduce the number of inter-lingual hyponymy links by nearly a half. What is more, the pilot stage of the works has also highlighted a number of cases which need to be carefully attended if the procedure is to be developed. These issues concern mainly the technical aspect of WordNets' design. In addition, from a theoretical linguistic perspective the described methodology of expanding the resources of Princeton WordNet leads to an analysis of its resources and relations holding between it and plWordNet that in turn allow multilingual aligned WordNets to become tools allowing for identification of lexical gaps and lexical areas that could be of interest in future semantic studies.

The advantages of the proposed strategy is certainly locating new PWN synsets in the proper nodes of the PWN relation structure and not changing the original structure of Princeton WordNet. On the other hand, there is definitely a risk for the Polish orientation of a constructed resource. In order to overcome it, lexicographers consult a variety of English sources in the course of their work.

6. Appendix

Procedures for lexicographers:

Group 1 synsets:

1. Check whether the generated English equivalent is correct.
2. If only one English equivalent was generated or if only one of the generated English equivalents is correct, then proceed to point 5. Otherwise, move to point 3.
3. If more than English equivalent was generated and more than one can function as the target for the plWN synset, then on the basis of the frequency lists the equivalent with the highest frequency and entropy is selected.
4. If none of the generated English equivalents is correct, then the lexicographers search for the matching equivalent and confront its occurrence with the frequency list
5. Introduce the new synset into PWN, connect it with a respectful hypernym or holonym, provide definition and usage examples for it
6. Choose the appropriate inter-lingual synonym for the English synset

Group 2 synsets:

1. Using your linguistic knowledge, bi-and monolingual dictionaries search for the matching equivalent and confront its occurrence with the frequency list
2. Introduce the new synset into PWN, connect it with a respectful hypernym or holonym, provide definition and usage examples for it
3. Choose the appropriate inter-lingual synonym for the English synset

Group 3 synsets:

1. Check the meaning of the PWN synset, if the PWN synset is a valid equivalent, then check whether the lack of inter-lingual synonymy link is not the mapping error, which should be immediately corrected
2. If the lack of inter-lingual synonymy is not the result of mapping error, then notify the supervisor about the problem with the synset and classify the problem into one of the following categories:

   [diff. syns] — the correct equivalent is located in a different synset

   [good. rel] — the established inter-lingual relation is correct

   [network.mismatch] — the lack of inter-lingual synonymy results from the differences in the relations networks in plWN and PWN
3. If the PWN synset is a not valid equivalent, then using your linguistic knowledge, bi-and monolingual dictionaries search for the matching equivalent and confront its occurrence with the frequency list
4. Introduce the new synset into PWN, connect it with a respectful hypernym or holonym, provide definition and usage examples for it
5. Choose the appropriate inter-lingual synonym for the English synset

References

Fellbaum, C. (1998). *Wordnet: An electronic lexical database.* Cambridge, Mass.: MIT Press.

Hamp, B. & Feldweg, H. (1997). Germanet: A lexical-semantic net for German. In *Proceedings of the Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.*

Kędzia, P., Piasecki, M., & Przybycień, E. R. K. (2013). Automatic prompt system in the process of mapping plWordNet on Princeton WordNet. *Cognitive Studies | Études cognitives*, (13), 123– 141.

Maziarz, M., Piasecki, M., Rudnicka, E., & Szpakowicz, S. (2013). Beyond the transfer-and-merge wordnet construction: plWordNet and a comparison with Wordnet. In *Proc. RANLP* (pp. 443– 452).

Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). Approaching plWordNet 2.0. In *Proceedings of the 6th Global Wordnet Conference.*

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*(4), 235– 244. http://doi.org/10.1093/ijl/3.4.235

Pianta, E., Bentivogli, L., & Girardi, C. (2002). Multiwordnet: Developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Piasecki, M., Maziarz, M., Szpakowicz, S., & Rudnicka, E. (2014). PlWordNet as the cornerstone of a toolkit of lexico-semantic resources. In *Proc. 7th International Global Wordnet Conference* (pp. 304– 312).

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A wordnet from the ground up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Piotrowski, T. & Saloni, Z. (2002). *Słownik angielsko-polski i polsko-angielski*. Warszawa: Wydawnictwo Naukowe PWN.

Pushpak, B. (2010). Indowordnet. In *Lexical Resources Engineering Conference 2010 (LREC 2010)*.

Rudnicka, E., Maziarz, M., Piasecki, M., & Szpakowicz, S. (2012). A strategy of mapping polish WordNet on Princeton Wordnet. In M. Maziarz, M. Piasecki, E. Rudnicka, & S. Szpakowicz (Eds.), *Proceedings of COLING*.

Svensen, B. (2009). *A handbook of lexicography: The theory and practice dictionary-making*. Cambridge: Cambridge University Press.

Vossen, P. (2002). *EuroWordNet General Document*.

Acknowledgment

This work was supported by the Polish Ministry of Science and Higher Education, project CLARIN-PL.

The authors declare that they have no competing interests.

The authors' contribution was as follows: ER and WW were responsible for the idea of the research, design of the procedures that served as the input for the algorithms and interpretation of the results. ER and WW drafted and wrote the manuscript. MK was responsible for retrieving the data from Słowosieć and WordNet databases, and for providing the quantitative information on the data and application of the algorithms. MK was also responsible for implementing the algorithms on the basis of procedures created by ER and WW.

Publisher: Institute of Slavic Studies, PAS, University of Silesia & The Slavic Foundation