Towards a methodology for filtering out gaps and mismatches across wordnets: the case of noun synsets in plWordNet and Princeton WordNet

Ewa Rudnicka
Wrocław University of Technology, Poland
ewa.rudnicka@pwr.edu.pl

Wojciech Witkowski
University of Wrocław, Poland
wojciech.witkowski@uwr.edu.pl

Łukasz Grabowski
Opole University, Poland
lukasz@uni.opole.pl

Abstract

This paper presents the results of large-scale noun synset mapping between plWordNet, the wordnet of Polish, and Princeton WordNet, the wordnet of English, which have shown high predominance of inter-lingual hyponymy relation over inter-synonymy relation. Two main sources of such effect are identified in the paper: differences in the methodologies of construction of plWN and PWN and cross-linguistic differences in lexicalization of concepts and grammatical categories between English and Polish. Next, we propose a typology of specific gaps and mismatches across wordnets and a rule-based system of filters developed specifically to scan all *I(inter-lingual)-hyponymy* links between plWN and PWN. The proposed system, it should be stressed, also enables one to pinpoint the frequencies of the identified gaps and mismatches.

1 Introduction

Since the development of the first wordnet, that is, Princeton WordNet (henceforth PWN, cf. Fellbaum, 1998), a number of wordnets for the multitude of languages have followed. Their construction was usually based on either of the two major approaches: the *merge* approach assuming manual wordnet creation on the basis of language data collected from dictionaries (e.g. Hindi Wordnet cf. Narayan et al 2001) and the *expansion* approach taking the content and structure of one of the existing wordnets as input for translation to another language (e.g. IndoWordnet, Bhattacharyya, 2010). Some wordnets were also built by means of the 'mixed', *transfer-and-merge* (also called *merge-expand*) method joining the previously mentioned approaches (cf. EuroWordNet, Vossen, 2002; Romanian Wordnet, Cristea et al, 2004; Open Multilingual WordNet, Bond and Foster, 2013). Thus, the process of their construction was often intertwined with the process of their linking to PWN, which served as the 'input' wordnet. The obvious advantage of the *expansion* and, partly, *transfer-and-merge* method is time and cost effectiveness, yet it looses on capturing the actual structure and content of the lexical system of a language in question. One of the few wordnets created independently of PWN is plWordNet, a wordnet of Polish language (henceforth plWN), built manually with the help of a unique method of extracting information on lexico-semantic relations from large text corpora (cf. Piasecki et al., 2009; Maziarz et al., 2014). Although much more time-consuming and expensive, such method of construction yields a resource which more closely reflects a lexical system of a language. The noun part of plWN has been already linked to PWN using a set of 7 inter-lingual relations (modelled on by those used in EuroWordNet, cf. Vossen, 2002). All of them were introduced manually by a team of bilingual lexicographers working in accordance with a detailed, three-stage mapping procedure (cf. Rudnicka et al., 2012). Already the first effects of this mapping process have showed a variety of contrasts in the structure and content of plWN and PWN. Some of them could be traced to different concept and (partly) grammatical categories' lexicalization between English and Polish; other resulted from different construction methods of plWN and PWN. The structure of Princeton WordNet was motivated by the results of psycholinguistic studies, while its content was largely based on individual lexicographers' choices and the data obtained from monolingual dictionaries.

In the paper, we present the results of a final stage of plWN to PWN noun synsets mapping and a proposal of a rule-based system of filters that enables one to identify the sources and measure the degree of gaps and mismatches between the two wordnets. The paper is organised as follows: in Section 2 the manual mapping strategy is described and the statistics of inter-lingual relations are given, in Section 3 different types of gaps and mismatches revealed in the course of mapping are discussed, in Section 4 a procedure for filtering out gaps and mismatches across wordnets is presented, in Section 5 the results of filtering are presented, while in Section 6 the conclusions are given.

2 Mapping results

Mapping between plWN and PWN was carried out by a team of trained and supervised bilingual lexicographers working in accordance with a detailed mapping procedure (cf. Rudnicka et al., 2012). The mapping was performed on the level of synsets (as in the case of all world wordnets) and consisted in linking plWN and PWN synsets corresponding in meaning and position in wordnet structure by means of one of 7 hierarchically ordered inter-lingual relations, such as S*ynonymy, Partial synonymy, Inter-register synonymy, Hyponymy, Hypernymy, Holonymy* and *Meronymy*. The mapping procedure consisted of three major steps: recognising the sense of the source synset, searching for the most corresponding target synset and selecting an inter-lingual relation to be established. In their work, lexicographers consulted the whole variety of available dictionaries and encyclopedias. Also, they were supported by a custom-designed system of automatic prompts, based on the relaxation labeling algorithm paired with a filtering by a large cascade dictionary (cf. Kędzia et al., 2013).

So far, the process of mapping has been conducted for noun and adjective synsets. The noun part is almost finished, the work on adjective part is still in progress. Therefore, in this paper we focus on the results of noun mapping. In Table 1, we compare basic numbers for plWN and PWN, while in Table 2 the counts of the established I(nter-lingual) relations are given.

Data analyzed	**plWN**	**PWN**	**plWN - Nouns**	**PWN - Nouns**
no. of synsets	198029	109505	123709	87695
no. of lexical units	269347	190049	166938	154385
no. of lemmas	182374	151162	126482	124879

Table 1: plWN 2.3. and PWN 3.1. general statistics[1]

I-relation	**Instances all**	**Instances nouns**
Synonymy	37191	33613
Hyponymy	85338	67680
Meronymy	6428	6428
Partial synonymy	5166	3767
Hypernymy	4142	4077
Holonymy	3025	3025
TOTAL	**141290**	**118770**

Table 2: plWN 2.3. to PWN 3.1. mapping statistics: instances of I-relations

One may plausibly argue that the most striking feature of the obtained results is the frequency of *I(inter-lingual)-hyponymy* links, which is two times higher than the frequency of the 'highest priority' *I(nter-lingual)-synonymy* links. Such results definitely point to a number of discrepancies between the content and structure of the two wordnets. Some sources of those discrepancies were already identified (Rudnicka et al., 2012): they encompass those due to the differences between lexical systems of English and Polish and those relating to different construction methods of the two wordnets under scrutiny. Still, the paper presents and discusses the results of only the very first stage of the mapping process. As shown in Table 2, the tendency of the double predominance of *I-hyponymy* over *I-synonymy* has prevailed and there arises the need to explain the reasons behind it.

[1] The data given in Table 1 and Table 2 are taken from the official plWordNet's website: http://plwordnet.pwr.wroc.pl/wordnet/stats

3 Gaps and mismatches across wordnets

In this section, we will discuss, first, the contrasts resulting from different construction methods of plWN and PWN and, second, various types of gaps and mismatches that may occur between lexical systems of natural languages (on the example of English and Polish).

Hence, the main research problem addressed in this paper refers to identification of any gaps and mismatches between linguistic data stored in two electronic lexical databases, that is, PWN and plWN. In general terms, the language-pair specific gaps and mismatches, which will be described in greater detail later in this paper, result from the following factors: 1) differences in structures of PWN and plWN; 2) differences in methodologies used to compile PWN and plWN; 3) specificity of mapping procedure applied to plWN and PWN; 4) systemic differences between English and Polish lexicon, morphology and syntax (e.g. varying degrees of lexicalization; differences in encoding of grammatical categories (e.g. gender); varying degrees of morphological productivity of affixal derivation); 5) cross-cultural differences between English and Polish. These differences, as applicable to the lexical data stored in plWN and PWN, are discussed in greater detail in the following sections 3.1 and 3.2.

3.1 Structural and methodological differences between plWN and PWN

In their analysis of the first stage of mapping of noun synsets, Rudnicka et al. (2012) identify two main sources of the observed predominance of I-hyponymy over I-synonymy links: these include contrasts in wordnet structure resulting from the application of different construction methods for each wordnet as well as morpho-lexico-semantic gaps and mismatches attributable to cross-linguistic differences between English and Polish. The former ones include the use of *Hyponymy and* vs. *Hyponymy or*, the use of different intra-lingual relations (*Hyponymy* vs. *Meronymy*) to capture the same conceptual dependencies and the occasional placement of mass/count, singular/plural and hyponym/hypernym lexical units in the same synset on the PWN side. The latter ones consist of greater degree of lexicalization of such grammatical categories in Polish as gender, diminutiveness and augmentativeness.

In the present study, the results of the final stage of mapping of noun synsets are analysed with an eye to other sources of the predominance of *I-hyponymy* relation over *I-synonymy* relation. Since mapping was carried from plWN to PWN side, we have searched for peculiarities of plWN's structure in order to develop a methodology that would lead to the creation of a large number of synsets lacking direct equivalents on PWN side. Three such groups of synsets were identified: gerund forms, multi-word expressions and forms belonging to marked registers. The latter ones will be discussed in the subsequent section, since most of them belong to the category of the so-called referential gaps (cf. Svensen 2009, also called 'cultural mismatches' cf. Bond et al. 2014). In plWN, there is a number of gerund forms under the category of noun. This is motivated by their ability to function as both verb participles and nouns. The creators of PWN did not adopt a similar strategy, hence there are not that many "-ing forms" in PWN noun synsets. Consequently, there could not be many *I-synonymy* links established in this category. The creators of plWN originally introduced many multi-word expressions and only recently a complex procedure for identifying multi-word lexical units has been applied (cf. Maziarz et al., 2015). The structural and methodological differences between PWN and plWN are summarized in Table 3 below:

plWN	**PWN**
hyponymy *and* {musical 1} - 'musical' hypo > {film 1}'film' {musical 2} - 'musical' hypo > {przedstawienie 7}- 'play'	**hyponymy *or*:** {musical 1} hypo > {movie 1}, hypo > {film 2}(a play or film whose action and dialogue is interspersed with singing and dancing)
use of different intra-lingual relations (hyponymy vs. meronymy) to capture the same conceptual dependencies	
{naszyjnik 1}[necklace] - mero-> {biżuteria 1} [jewellery]	{bracelet 2} hyponymy > {jewellery 1}
mass and count nouns in the same synset	
{mebel 1} (piece of furniture),{umeblowanie 2}(furniture)	{furniture 1, piece of furniture 1} 'furnishings that make a room or other area ready for occupancy'
singular and plural in the same synset	
{pieróg 2} 'small	{dumpling 1, dumplings

boiled ball of dough with various stuffing’	1} ‘small balls or strips of boiled or steamed dough
gerunds in plWN	
{kopanie 2} ‘the act of kicking’	------------------------------
plWN multi-word synsets	
{eskadra bobmowa 1} ‘bomber squadron’ {eskadra niszczycieli 1} ‘destroyer squadron’	------------------------------ ------------------------------

Table 3: Structural and methodological differences between plWN and PWN handled by *I-hyponymy* relation

3.2 Morpho-lexical mismatches ad lexico-semantic gaps

As already mentioned in the previous section, the second important source of the high frequency of *I-hyponymy* links between plWordNet and Princeton WordNet identified by Rudnicka et al. (2012) are the differences between lexicalisation (and structuralisation) of concepts and grammatical categories between English and Polish. The latter ones are called **morpho-(syntactic) mismatches** by (Bond et al., 2014: 252). They result from systemic differences between languages; in practice this means varying degrees of lexicalization of certain grammatical categories, such as number or gender (e.g. Pol. *kuzyn/kuzynka* vs. Eng. *cousin*; Pol. *Amerykanka* vs. Eng. *American girl*). In other words, certain concepts are “lexicalized through words with different morpho-syntactic properties across languages” (ibid.). [This resembles what Catford (1965/1978) refers to as category shifts in the context of translation]. Such differences may also result from varying degrees of morphological productivity of derivational morphemes, notably in the case of diminutives (e.g. Pol. *samochód / samochodzik* vs. Eng. *car*), augmentatives (Pol. *dom/domisko* vs. Eng. *house*). Due to its productivity, we expect a high number of such cases in plWN to PWN mapping. Also, their recognition should not pose major problems, since they can be identified by intralingual plWN morpho-lexical relation links holding between lexical units, such as *Żeńskość* - ‘Feminine gender’, *Diminutywność* - ‘Diminutiveness’ and *Augmentatywność* - ‘Augmentativeness’ (cf. Maziarz et al., 2012).

The more challenging part are the differences arising from different lexicalisation of concepts. These are widely discussed in the literature. Cvilikaite (2006: 129) argues that the so-called *lexical gaps* should be identified on the level of individual meanings of lexical items. The reason for that is that translators are usually interested in context-specific individual meanings of lexical items rather than semantic structures of lexemes, often polysemous ones (ibid.). In fact, lexical gaps occur when a given concept is not lexicalized in a given language (Cvilikaite, 2006) or when it is it is expressed with a lexical unit in one language and with a free combination of words in another language (Bentivogli, Pianta and Pianesi, 2000; Hutchins & Somers, 1992). In specialist literature, one may find a number of typologies of lexical gaps and mismatches between data stored in bilingual dictionaries or multilingual wordnets (e.g. Svensen, 2009; Bond et al., 2014); also, specialist literature on translation studies and linguistic typology addresses the problem of incompatibility of lexicons of different languages (e.g. Talmy, 2000). In this paper, we aim to synthesize the aforementioned typologies in order to capture lexical gaps and mismatches between linguistic data, more specifically, between nouns stored in PWN and plWN.

The first group are **referential gaps** (Svensen 2009: 271), which roughly correspond to what Bond et al. (2014: 252) subsume under an umbrella label of ‘cultural concepts’. These include culture-specific concepts that are lexicalized in one language and not lexicalized in another. Such concepts are tied to the history, customs, traditions making up the cultural heritage of a given linguistic community. For example, concepts such as *szmalcownik* ‘a person who extorted money from Jews under threat of denouncing on them; a word used in the period of German occupation of Poland during World War II’ or *noc Kupały* or *kupała* ‘summer solstice celebrated on the night of 23/24 June, the shortest night during entire year’ are cultural concepts specific to or deeply rooted in the Polish culture and hence not lexicalized in English. In a similar vein, names of national holidays, institutions, administrative functions and units, historical names, etc. fall into this category.

The next group are the so-called ‘pragmatic lexicalizations’ (Bond et al., 2014: 252), which correspond to what Svensen (2009: 273) refers to as **lexical gaps**. In short, these include concepts that are familiar in many cultures yet they are not lexicalized in all of them (Bond et al., 2014: 252). Because such concepts are known across cultures, they reveal differences in lexicalization of their conceptual structure e.g. Eng. *uncle* vs. Pol. *stryj/wuj*; Pol. *palec* vs. Eng. *finger/toe*. The last group of gaps resulting from cross-cultural differences are the so-called **differences in perspective** (Bond et al., 2014: 252) or **standpoint gaps**, that is, the differences resulting from structuring conceptual reality from various perspectives or standpoints (who does what to whom and how) e.g. Eng. *married* vs. Pol. *żonaty/mężatka*; Eng. *house/home* vs. Pol. *dom*; Eng. *bring/take* vs. Pol. *przynieść*. Table 4 summarizes the gaps and mismatches discussed in the foregoing.

plWN	**PWN**
Differences arising from productive morphological derivation	

Diminutives:
{samochód 1} 'a car',
{samochodzik 2} 'a small car'
Augmentatives:
{dom 1} 'a house',
{domisko 1} 'a large house'
	Diminutives:
	{car 1}
	Augmentatives:
	{house 1}

Referential gaps/Cultural concepts	
{szmalcownik 1} 'blackmailer' {noc Kupały 1} 'summer solstice celebration'	--------------------------------- ---------------------------------
Lexical gaps/Pragmatic lexicalization	
{stryj 1} 'father's brother',{wuj 1} 'mother's brother' {palec 1} 'digit of a hand or foot' {kończyna górna 1} 'upper limb', {kończyna dolna 1} 'lower limb'	{uncle 1} 'the brother of your father or mother' {finger 1}, {toe 1} {limb 1} 'one of the jointed appendages of an animal used for locomotion or grasping'
Differences in perspective/ Standpoint gaps	
{żonaty 1} 'married man', {mężatka 1} 'married woman'	{married 1} 'a person who is married'
Morpho-lexical mismatches: grammatical gender lexicalization	
{kuzyn 1} 'male child of your uncle or aunt' {kuzynka 1} 'female child of your uncle or aunt'	{cousin 1} 'the child of your aunt or uncle'

Table 4: Taxonomy of gaps and mismatches between plWN and PWN

4 Methodology: a procedure for filtering out gaps and mismatches

In this section, we propose a rule-based system of filters designed for the recognition of the different types of gaps and mismatches that may occur in wordnet mapping. Based on the typology of gaps and mismatches described in Section 3, the system scans all *I-hyponymy* links from plWN to PWN side. Its ultimate aim is to filter out, first, contrasts resulting from different construction methods of plWN and PWN, second, all and any systematic mismatches resulting from different lexicalization patterns of grammatical categories, third, cultural gaps. Ultimately, the system aims to produce the set of proper lexical gaps. The system's implementation is conducted in a number of steps presented in greater detail below.

Step 1. I-hyponymy

- select all plWN **noun** synsets that have I-hyponymy relation to PWN synsets.
- Create a list of plWN - PWN noun synset pairs.

Step 2. From the list obtained in [1] filter out:

- all plWN gerund forms. Do this by filtering out those synsets whose L(exical)U(nit)s have *Synonimia międzyparadygmatyczna V-N (Cross-paradigm Verb-Noun synonymy)* relation
- all plWN synsets that belong to [sys(tematics)] domain
- all plWN synsets built from LUs denoting proper names or LUs derived from proper names. Do this by removing all plWN synsets which have Typ/*Egzemplarz (Type / Instance)* relation.
- all plWN multi-word synsets which are not tagged as multi-word (fixed) phrases in plWN
- (keep on a separate list) all Princeton WordNet synsets that are built in the following manner: {LU1 (lemma1)}, {LU2 (lemma1+s)} or {LU1 (lemma1)}, {LU2 (lemma1+ing)})

Step 3. Filtering out morpho-lexico-syntactic mismatches. From the set remaining after completion of [Step 2], sort out all plWN synsets that include lexical units which have specific intra-lingual lexical unit relations (such as (1st) *żeńskość (feminine form),* (2nd) *diminutywność (diminutiveness) & augumentatywność (augumentativeness)*) to other plWN LUs. For each filtering stage save the list of filtered out results.

Step 4. From the set remaining after [Step 3] has been carried out, (tests for filtering out cultural gaps) - remove PWN synsets with relation *Topic/Domain*[2] (keep on a separate list)

Step 5. Filter out Polish domain specific synsets - From the list of synsets remaining after the implementation of [Step 4], sort out synsets containing LUs belonging strictly to Polish language domain. The target are those synsets whose LUs have the following register markers[3]: ##K: pot., ##K: posp., ##K: wulg., ##K: daw., ##K: środ. or ##K: reg. marked.

5 Results

The results are summarized in Table 5. The filtering procedure resulted in removing out only 44.83% i.e. 30679 synsets out of the overall 67680 plWN synsets mapped onto PWN synsets by means of *I-hyponymy* relation. The biggest percentage of those constitute gerund forms and proper names (21%). The former ones, together with multi-word synsets[4] (5.39%) (both removed in Filter 2), are the effect of plWN's methods of construction. The next groups in line are diminutives and augmentatives (5.32%) and feminine forms (3.73%) (removed in Filter 3) which reflect the specificity of Polish morphology. Another group are PWN synsets that have the intralingual *Topic-Domain* relation (5.79%), removed in Filter 4 aimed at removing mainly culture-dependent concepts found in PWN. The last and the least numerous group are plWN synsets including lexical units marked for register (3.4%), also aimed at removing culture-specific concepts.

With respect to the data presented in Table 5, it should also be noted that the remaining number of 37001 synsets is too large for any manual analysis and hence it needs to be treated with caution; the said number is primarily influenced by the size differences between plWN and PWN. Accordingly, in order to minimize the effect of database size, the results of filtering were divided into three groups defined in terms of dictionary and wordnet coverage. The results of this division are presented in Table 6.

F	Details	no. of synsets removed	% removed
2	gerunds, pr. names, [sys] domain	14478	21%
	plural number errors	155	0.2%
	multi-word synsets (but not multi-word units)	3649	5.39%
3	diminutives and augmentatives	3606	5.32%
	feminine form	2526	3.73%
4	topic / domain relation	3921	5.79%
5	[posp] - everyday common	91	0.13%
	[pot] - everyday non-standard	1137	1.68%
	[reg] - regional variants	173	0.2%
	[srod] - social group specific	0	0%
	[wulg] - vulgar	9	0.01%
	[daw] - archaisms	934	1.38%
TOTAL REMOVED		30679	44.83%
REMAINING - candi-		37001	-------------

[2] In [Filter 3] all plWN synsets that hold *I-hyponymy* relation to PWN synsets with *Topic/Domain* relation within PWN are removed. That may seem a 'drastic' move, yet we aimed at removing all potential cultural gaps. Thus, the number of synsets removed by [Filter 3] - 3921 - should be treated with caution as it is overestimated, since it also includes synsets that lexicalize concepts common to both Polish and English.

[3] The abbreviations used to mark relevant registers are explained in Table 5.

[4] What is meant by a multi-word synset is a synset whose LUs are built of more than one word but are not treated as multi-word units in the sense of Maziarz et al. (2015) e.g. {eskadra niszczycieli 1} - 'fighter sqaudron', where multi-word units are defined as those composed of a sequence of words that cannot be separated from each other and occur in a fixed order, e.g. {chlorek amonu 1} - 'ammoniun chloride'.

dates for actual lexical gaps		

Table 5: Filtering procedure results

Synset type	**Instances**
[Group 1] - synsets whose lemma are not present in Princeton WordNet and whose equivalent was found in a 'cascade dictionary'	9420
[Group 2] - synsets whose lemma are not present in Princeton WordNet and whose equivalent was not found in a 'cascade dictionary'	18567
[Group 3] - synsets whose lemma are present in Princeton WordNet and whose equivalent was found in a 'cascade dictionary' but which are not related via I-synonymy Pol-Eng or I-partial synonymy Pol-Eng relation	9014

Table 6: Group division of possible candidates for lexical gaps in plWN and PWN comparison

The data in Table 6 show that the number of candidates for actual lexical gaps can be lowered by 9420 Group 1 synsets, a decision that yields 27581 possible candidates. The resulting number, however, is still too large for a manual analysis. However, with due caution it can be lowered by the reduction of cases in Group 2, where, given large enough language resources, a substantial number of English equivalents can be found.

6 Discussion and conclusions

This study constituted an attempt at identifying any gaps and mismatches between lexical data, specifically, nouns, stored in two inter-liked electronic databases, that is, plWN, the wordnet of Polish, and PWN, the Princeton wordnet of English. The results confirmed our initial hypotheses that the gaps and mismatches result from wordnet-specific structural and methodological differences, specificity of the interlingual mapping procedure, systemic differences between Polish and English as well as cross-cultural differences. In order to identify any gaps and mismatches across two aforementioned wordnets, a custom-designed filtering procedure was developed and described in this paper. The results of filtering procedure revealed groups of synsets (mainly gerunds and multi-word synsets) whose existence and high numbers are the effect of the assumed methodology of construction of plWN. Next, the procedure revealed groups of synsets such as diminutives, augmentatives and feminine forms that reflect the specificity of the morphology of the Polish language and, finally, PWN synsets holding the relation *Topic-Domain* and plWN synsets marked for different registers that attest to the presence of culture-dependent concepts were filtered out/identified.

The approach presented in this study has a number of limitations which need to be addressed in future research. First, the results revealed up to 28000 synsets that are required to be manually analyzed, the process that is bound to be time-consuming and labour-intensive. Second, the procedure described in this paper does not allow, in its current form, checking the filtering results against larger lexical resources (e.g. larger than the cascade dictionary used in this study) where more potential equivalents for Polish lemmas could be found. To this end, it is possible to use additional resources such as Polish-English parallel corpora (e.g. PARALELA[5], a large collection of Polish-English parallel texts); this could provide an improvement in terms of filtering out the results. A small-scale manually conducted experiment aimed at identification of equivalents in corpora and Internet resources has revealed that the number of lemmas present in plWN and, at the same time, not found in the cascade dictionary used in the filtering procedure can be lowered by approximately 37% (see Rudnicka and Witkowski, 2015). Finally, it should be stressed that at the current stage there are no means of filtering out exactly those Polish synsets whose potential equivalents could be multi-word units with compositional meaning in English. Removal of all plWN multi-word synsets with no special tag (cf. Maziarz et al., 2015 for separate treatment of multi-word units in plWN) as a part of [Filter 2] appears to be a significantly imprecise tool with respect to compositionality of meaning, i.e. this operation removed all multi-word synsets in one fell swoop, regardless of the internal semantic dependencies of the words in a multi-word unit. To resolve this problem, it is possible to target the relevant plWN multi-word synsets by identifying instances in which the synsets at hand have *Hyponymy* and *Meronymy: element* relation links to

[5] http://paralela.clarin-pl.eu

synsets whose LUs have the same bases as the LUs in question.

As for the future, the procedure described in this study and its results may come in useful for exploration of the different types of equivalence relations obtained between lexical data stored in plWN and PWN. This could enable one to turn the study results into actionable knowledge useful for lexicographers and translators, among others.

Acknowledgement

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Arleta Adamska-Salaciak. 2014. Bilingual Lexicography: Translation Dictionaries. *International Handbook of Modern Lexis and Lexicography.* Springer-Verlag, Berlin Heildelberg.

Luisa Bentivogli, Emanuele Pianta and Fabio Pianesi. 2000. Coping with lexical gaps when building aligned multilingual wordnets. [In:] *Proceedings of LREC 2000*, Athens, Greece, pp. 993-997.

Pushpak Bhattacharyya. 2010. IndoWordNet. Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May, 2010.

Francis Bond. 2005. Translating the Untranslatable. A Solution to the Problem of Generating English Determiners. *CSLI Studies in Computational Linguistics*.

Francis Bond, Christiane Fellbaum, Shu-Kai Hsieh, Chu-Ren Huang, Adam Pease and Piek Vossen 2014. A Multilingual Lexico-Semantic Database and Ontology. [In:] *Towards the Multilingual Semantic Web* Paul Buitelaar and Philipp Cimiano (eds), Springer pp 243–258 (Publisher's page). http://compling.hss.ntu.edu.sg/who/bond/pdf/2014-msw-omw.pdf

Jurgita Cvilikaite. 2006. Lexical Gaps. Resolution by functionally complete units of translation. *Darbai ir Dienos*, 45, 127-142. donelaitis.vdu.lt/publikacijos/dd45_cvilikaite.pdf

Christiane Fellbaum (ed). 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusets.

W. John Hutchins and Harold L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press, London.

Paweł Kędzia, Maciej Piasecki, Ewa Rudnicka and Konrad Przybycień. 2013. Automatic Prompt System in the Process of Mapping plWordNet on Princeton WordNet. *Cognitive Studies,* 13: 123-142.

Dipak Kumar Narayan, Debasri Chakrabarty, Prabhakar Pande, Pushpak Bhattacharyya. 2001. *An Experience in Building the Indo WordNet - a WordNet for Hindi*. 1st International Conference on Global WordNet (GWC 02), Mysore, India.

Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. Approaching *plWordNet* 2.0. [In:] *Proceedings of the 6th Global Wordnet Conference*, Matsue. 189-196.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2013. Beyond the Transfer-and-Merge WordNet Construction: plWordNet and a Comparison with WordNet. [In:] *Proceedings of RANLP*, Hissar.

Marek Maziarz, Stan Szapkowicz and Maciej Paisecki. 2015. A Procedural Definition of Multiword Lexical Units. [In:] *Proceedings of RANLP, Hissar.*

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda 2009. *A Wordnet from the Ground Up.* Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. [In:] Proceedings of COLING 2012. ACL.

Ewa Rudnicka and Wojciech Witkowski. 2015. Towards the Methodology for Extending Princeton WordNet. *Cognitive Studies 15.*

Bo Svensen. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making.* Cambridge University Press, Cambridge.

Leonard Talmy. 2000. *Toward a Cognitive Semantics. Typology and Process in Concept Structuring*. MIT Press, Cambridge, Massachusetts.

Piek Vossen (ed.). 2002. *EuroWordNet General Document,* Version 3 (final) URL: http://www.hum.uva.nl./~ewnAlfred.