Lodz Papers in Pragmatics 13.1 (2017): 3–24
DOI:10.1515/lpp-2017-0002

Ewa Rudnicka
Wrocław University of Technology

Francis Bond
Nanyang Technological University

Łukasz Grabowski
University of Opole

Maciej Piasecki
Wrocław University of Technology

Tadeusz Piotrowski
University of Wrocław

TOWARDS EQUIVALENCE LINKS BETWEEN SENSES IN PLWORDNET AND PRINCETON WORDNET

Abstract
The paper focuses on the issue of creating equivalence links in the domain of bilingual computational lexicography. The existing interlingual links between plWordNet and Princeton WordNet synsets (sets of synonymous lexical units – lemma and sense pairs) are re-analysed from the perspective of equivalence types as defined in traditional lexicography and translation. Special attention is paid to cognitive and translational equivalents. A proposal of mapping lexical units is presented. Three types of links are defined: super-strong equivalence, strong equivalence and weak implied equivalence. The strong equivalences have a common set of formal, semantic and usage features, with some of their values slightly loosened for strong equivalence. These will be introduced manually by trained lexicographers. The sense-mapping will partly draw on the results of the existing synset mapping. The lexicographers will analyse lists of pairs of synsets linked by interlingual relations such as synonymy, partial synonymy, hyponymy and hypernymy. They will also consult bilingual dictionaries and check translation probabilities in a parallel corpus. The results of the proposed mapping have great application potential in the area of natural language processing, translation and language learning.

Keywords
equivalence, wordnets, interlingual mapping, synset, lexical unit, sense-level

1 Introduction

The paper presents a proposal of a system of equivalence links between lexical units of plWordNet of Polish (cf. Piasecki et al 2017) and Princeton WordNet of English (cf. Fellbaum 1998). Wordnets are big electronic lexico-semantic databases in which words, or more specifically, their *lexical meanings* (also called *word senses*), are connected by a rich network of lexico-semantic relations such as synonymy, hyponymy, or antonymy, to name just a few. The relations are the main determinant of meaning in a wordnet, although they are supplemented by *glosses*[1] and *usage examples*. For each sense, also a coarse-grained *semantic domain* is given (called a *lexicographic file* in Princeton WordNet). Moreover, in plWordNet the information about the *stylistic register* of each lexical unit is provided (whether it is general or any of the marked ones). Many wordnets are already inter-connected forming large, multilingual networks, the biggest one being Open Multilingual WordNet (cf. Bond and Foster 2013). There are two basic units of organisation within a wordnet: *lexical units* that are lemma – Part of Speech – sense triplets and *synsets* that can be shortly and informally described as sets of synonymous lexical units. The inter-linking between wordnets takes place at the level of synsets. This is unlike in traditional bilingual dictionaries that offer (a set of) equivalents for each of the senses of a selected word. Still, as in traditional dictionaries, the choice of the right equivalent (for a given context) is left to a dictionary or wordnet user.

For some time there has been a trend to help a dictionary user in this task by providing different types of clues and hints (cf. Crenn 1996; Yong and Peng 2007; Adamska-Sałaciak 2013; Lew 2013; Kamiński 2016). First, the specific sense of a word is signalled by its part of speech and often by its synonym in the source language. Sometimes also the name of a semantic domain and/or a register is given as a label. This information helps a dictionary user to identify the meaning of a source language item and select the right sense listed in a dictionary. However, substantially less help is offered in the choice of the best equivalent in the target language. Depending on the dictionary, usage examples are given; it also happens that source-language lexical item in a dictionary is embedded in a sentence (which is also translated) or additional minimal context is provided (e.g. with a collocate of a source-language item) and translated as a longer phrase (Adamska-Sałaciak 2013: 223). Such solutions are used to handle the problem of interlingual anisomorphism (that is, the fact that different languages structure reality in

[1] A gloss is a form of short definition of a lexical unit (a triple: lemma, Part of Speech, sense identifier, e.g. *zamek: N 1* 'a castle') that comments on the intended meaning targeted and supplements the primary definition in a form of the network of lexico-semantic relations. Glosses are introduced to facilitate applications of plWordNet in the natural language processing as well as human users comprehension of the meanings of the lexical units.

different ways) and a lack of absolute meaning equivalence, often called synonymy, between two languages (Adamska-Sałaciak 2013: 223–225). On the other hand, bilingual wordnets always provide source language item synonyms (if there are any), and a set of synonymous target language equivalents (if there are more than one). The number of synonyms and the amount of information going with them depends on methodological assumptions underlying the construction of a wordnet. These affect granularity of senses and the type and richness of semantic information.

In an ideal world, a dictionary or a wordnet user would be offered enough information to always properly identify the sense of the searched language item and choose its right equivalent. That would require a very fine-grained network of inter-lingual links (and a very rich network of intra-lingual links). Computers allow us to do both. In this paper, we will sketch a proposal of a strategy of linking wordnets at the level of lexical units, its smallest building parts. It will be based on the results of the existing synset-level mapping between plWordNet and Princeton WordNet and the recent theories of equivalence developed in bilingual lexicography literature (cf. Piotrowski 2011; Adamska-Sałaciak 2014).

2 Equivalence in bilingual dictionaries and wordnets

2.1 Equivalence types

What both bilingual dictionaries and bilingual wordnets do is to provide pairs, or sets of language items corresponding in, broadly speaking, meaning (called *equivalents*, esp. in dictionaries.). The status of the items and the degree of their correspondence may differ in both types of resources. It seems an obvious assumption that a lexical resource should aim at including only language items of a confirmed lexicality (cf. Svensen 2009: 102–103). Yet, in the absence of direct lexicalised equivalents, some dictionaries and some, especially translated, wordnets provide non-lexicalised multi-word target language expressions as equivalents. Some dictionaries mark them with a special font or font attribute to signal the lack of a lexicalised equivalent (cf. Svensen 2009). Some wordnets put them in the so called "artificial" synsets to mark their non-lexicality; MultiWordNet calls them *phrasets* (cf. Bentivogli and Pianta 2004; Finnet, Lindén and Carlson 2010). An example is the translation of the English synset {toilet_roll 1} to an Italian synset { GAP } (signalling a lexical gap in Italian) and next to an Italian phraset {rotolo_di_carta_igienica}. Some equivalence theorists, e.g. Piotrowski (2011) or Adamska-Sałaciak (2014:7), call them *explanatory* equivalents.

> Unlike a cognitive or a translational equivalent, an explanatory equivalent is not an established TL unit, but a free TL [Target Language] combination: a succinct paraphrase of the meaning of the SL [Source Language] headword. It resembles a mini-definition in the target language, except that it is normally short.

A different kind of strategy is to employ a TL word that can be used in translating a phrase or sentence, and which does not correspond on all features with the ST word, for example it has a different grammatical category; this does not often happen in bilingual dictionaries (such cases are handled in examples), but it does occasionally happen in wordnets. An example could be English *He is a good cook* corresponding to Polish *(On) dobrze gotuje*. Such equivalents, whose main feature is the absence of word-level correspondence (Héja 2016: 4), are called *functional* by Adamska-Sałaciak (2014: 7):

> functional equivalence is a relation holding not between the meanings of individual lexical items but between the meanings of longer stretches of text. Typically, the TL text portion either contains a TL word of a different grammatical category than the SL headword or features no element whatsoever directly corresponding to that headword.

Some bi- and multilingual wordnets, e.g. plWordNet mapped onto Princeton WordNet, employ an interlingual relation of cross-categorial synonymy to show correspondence in meaning between synsets belonging to different grammatical categories (cf. EuroWordNet, Vossen 2002; enWordnet, Rudnicka et al 2015.) This is done to at least partly specify the sense of the SL synset in addition to providing only a very general inter-wordnet hyponymy link, cf. the strategy employed in the mapping of adjectives between plWN and WN (Rudnicka et al 2015), for instance:

(1) {**neokatechumenalny** 1 `~related to neocatechumenate '}

inter-lingual-hyponymy->{*related* 1}
interlingual-cross-categorial-synonymy-> {*neocatechumenate* 1}

Still, there are no claims made with regard to the possibility of using inter-lingual cross-categorial synonyms as any kind of equivalents in translation, although such a possibility cannot be excluded, especially when lexical translation is combined with the adaptation of the target language syntactic-semantic structure of the translated expression.

Neither explanatory nor functional equivalents will be our main focus here. We instead focus on equivalence links only between lexicalised source and target language expressions. Lexicographers distinguish between two basic types of such equivalents, which can have different names *cognitive* (or semantic) and *translational*. We have to stress that the term *cognitive* has little relation to

"cognitive" as used in cognitive linguistics. It is a shorthand description of the fact that this relation is based on mental associations, and these are founded on "naive semantics", rather than another relation, *translational*, based on actual use in concrete translation. Adamska-Sałaciak (2014: 6) defines *cognitive equivalent* in the following way:

> A cognitive equivalent has a high explanatory potential, i.e., it is capable of faithfully rendering the meaning of the SL headword. Its identification is often more or less effortless, because it tends to spring to mind immediately after a bilingual speaker (lexicographer) has been presented with an SL headword. Thanks to that, cognitive equivalents are frequently identical in different dictionaries for the same language pair, which gives rise to the feeling that they are somehow "real" or "true." On the downside, a cognitive equivalent is often too general to work as a translation of the SL item in a particular context.

Clearly, the cognitive equivalent, which first comes to mind of a competent bilingual speaker, without him or her being presented with any context (Adamska-Sałaciak 2010: 400)[2], is the typical "naive meaning" equivalent, while not always the "use" equivalent. Some contexts require more specific and less general equivalents. Therefore, apart from cognitive equivalents, bilingual dictionaries also list translational equivalents, which are referred to as *contextual equivalents* by Héja (2016: 3); they are defined by Adamska-Sałaciak (2014: 7) as follows:

> A translational equivalent, by contrast, while not being wholly identical in meaning to the SL headword, produces a good translation when substituted for it in a particular context (not least because it has similar combinatory properties).

We have to note that the phrases that Adamska-Sałaciak uses, "to faithfully render the meaning" or "to be wholly identical in meaning", are based on the assumption that there is a simple method of comparing the meaning of two items from two languages, i.e., that meaning is finite and easy to describe. Unfortunately, that is not the case, but we shall not be concerned with this problem in this paper [see Taylor (2012) for a more detailed discussion].

---

[2] Adamska-Sałaciak (2010: 400) assumes that identification of a cognitive equivalent by skilled bilinguals "is characterized by a high-degree of intersubjective agreement". In dictionary-making this task is performed by lexicographers, who are skilled bilinguals (ibid.), and who use their linguistic intuition. While Adamska-Sałaciak does not refer to any empirical studies on the psychological foundations of lexicographers' work, she bases her observations on her experience in practical lexicography. In fact, we do not know any psycholinguistic work on lexicographer intuitions, and we also have our experience in practical lexicography.

To provide an example of the difference between cognitive and translational equivalents, a translational equivalent of the English adjective *heavy*[3] may be, depending on context, *wzmożony* or *duży* (e.g. *heavy traffic*) or *mocny* (e.g. *heavy make-up*) or even *pogrubiony* (e.g. (*to write in) heavy type*), yet its cognitive equivalent (*ciężki*), which first comes to mind of a naive user of both languages, is a very general one, and, while it may be successfully used in many contexts (e.g. *heavy bag, heavy suitcase, heavy fighting*), it may not fit some specific contexts. That is why Héja (2016: 7) argues that "cognitive equivalents and contextual [translational] equivalents should be considered as two ends of a scale", that is, from perfect to occasional to rare interchangeability, and that their position on the scale should be determined using a mathematical formula called conditional probability (ibid., 8). In other words, the position on the scale presents the potential of the use of equivalents in translation in a given context. From yet another perspective, Héja (2016: 7) notices that the scale also reflects the degree of symmetry between equivalents, namely that [perfect] cognitive equivalents are symmetric while translational equivalents tend to be asymmetric. Obviously, the degree of symmetry would depend on adopted criteria.

Adamska-Sałaciak (2014: 7) remarks further that since the number of contexts a SL item may appear in is unlimited, so is the number of its possible translations. For reasons of size limitations, dictionaries typically list only the most frequent ones (ibid.) For some, hints may be given with respect to their appropriateness in particular contexts, which will make them more useful for language learners and translators.

Clearly, the whole typology presented so far is more of interest to dictionary compilers or researchers rather than to dictionary users. The latter are not informed about the typological status of different equivalents for a given word sense. Often the differences between cognitive and translational equivalents are subtle and they are hard to distinguish even for a professional lexicographer. This is not the goal of constructing dictionaries. They aim to be of the greatest possible help for language users. On the other hand, bi- and multilingual wordnets have been and are constructed more as machine-readable dictionaries for natural language processing tasks (such as automatic translation, information retrieval etc.) than as resources for translators or language learners. Still, due to their large data coverage and richness in lexico-semantic information, they have been found a valuable resource for language users too. For example, plWordNet is accessed online over 10,000 times a month and linked to Princeton WordNet of English is used in the popular Polish multilingual online dictionary Ling.pl: <http://ling.pl>. The Japanese wordnet, also with the Princeton WordNet is used as as a bilingual resource in a

[3] This example is based on an entry in *PWN-Oxford Polish-English Dictionary*.

DOI:10.1515/lpp-2017-0002

variety of language learning sites, such as <http://www.manythings.org/japanese/reading/wnjpn/>.

2.2 Equivalence relations in wordnets

The majority of bilingual wordnets employ simple, one-to-one equivalence links between source and target synsets, with Princeton WordNet English synsets usually functioning as target ones (cf. EuroWordNet, Vossen 2002; Open Multilingual Wordnet, Bond and Foster 2013). The links are sometimes established manually (e.g. GermaNet cf. Hamp and Feldweg 1997), plWordNet, cf. Rudnicka et al 2012). In most cases the WN structure is copied and next semi-automatically filled with translations of synset members, e.g. in the case of SlowNet (the wordnet of Slovene, cf. Fišer and Sagot (2015)) this is done using multilingual aligned corpora. Occasionally WN synsets are only manually translated (e.g. FinnNet, the wordnet of Finnish, cf. Lindén and Carlson (2010). Such simple equivalence links are called inter-lingual synonymy by wordnet developers (starting from Vossen 2002). Similarly to bilingual dictionaries, they include cognitive and translational equivalents. However, since different wordnets are based on slightly different theoretical assumptions and, consequently, their synsets may differ in granularity (the subtlety of meaning distinctions resulting in the number of lexical units per synset), the simple equivalence link between two synsets does not entail that all lexical units from both synsets can function as each other's perfect translational equivalents.

(2) {**pies 2** - 'dog', **pies domowy 1** - 'domestic dog'} (zw) (semantic domain: 'animal')
pies 2 register: general;
gloss: *pies domowy - popularne zwierzę domowe, przyjaciel człowieka* - ' domestic dog - a popular pet, man's friend'
usage example: *Pies był bez kagańca*.
'The dog didn't have a muzzle.'

pies domowy 1 register: specialised; no gloss; no usage example given
Synset relations:
hyponym of {pies 1 - 'canine'} (zw) and {czworonóg 1 - 'quadruped'} (zw)
hypernym of {kundel 1 - 'mongrel'} (zw), {pies myśliwski 1- 'hunting dog'} (zw)

interlingual synonym of:
{**dog 1, domestic dog 1, Canis familiaris 1**} (zw) (semantic domain: 'animal')
gloss: a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds;
usage example: "the dog barked all night"

Looking at the members of synsets in example (2), a competent bilingual speaker (or a translator) will identify pairs of lexical units that are better and worse translational "matches". The better ones are clearly *pies* 2 and *dog* 1, and *pies domowy* 1 and *domestic dog* 1. Apart from the shared meaning, they also agree in register: general in the former pair, specialised in the latter and in the internal structure: simple word vs. multi-word expression. It has not been studied, as far as we know it, what those beliefs are based on, it is quite likely that their source is folk (naive) assumptions about languages, in European culture very strongly influenced by writing.

Such many-to-many links do not exist explicitly in bilingual dictionaries, though they can be recovered implicitly. What they usually offer implicitly is a one-to-many link, or one-to-one link. The former one appears and it is also quite frequent in wordnets when we deal with a single lexical unit source synset linked to a multiple lexical unit target synset, or the other way round. Such a case is illustrated in example (3) below:

(3) {**olej słonecznikowy 1** - 'sunflower oil'} (jedz) (semantic domain: 'food'),
register: general;
gloss: *olej roślinny, spożywczy wytwarzany z nasion rośliny oleistej - słonecznika*
'vegetable oil, foodstuff, produced from the seeds of oil plant - sunflower'
hyponym of {olej roślinny 1 - 'vegetable oil'}

interlingual synonym of:
{**sunflower oil 1, sunflower-seed oil 1**} (semantic domain: 'food'),
gloss: oil from sunflower seeds
hyponym of {vegetable oil 1}

Despite the fact that the English synset includes two lexical units: *sunflower oil* 1 and *sunflower-seed oil* 1, both seem to be very good translations of the Polish lexical unit *olej słonecznikowy* 1.

Intuitively, equivalence is perceived as a relation between *one* source language item and *one* target language item. However, it does not need to be so, as illustrated by example (3) above. It is also not such a frequent case in bilingual dictionaries or in bilingual wordnets. In the former, it occurs when only one equivalent is listed for a source language headword, in the latter when the equivalence link is established between two synsets including one lexical unit each. Below we provide an example (4) of the latter with the pair of plWN and WN synsets linked by an inter-wordnet relation of interlingual synonymy:

(4) {**indyk 1**} (jedz) 'turkey' (semantic domain: 'food'),
register: general; gloss: *mięso z indyka*- 'the meat from turkey',
usage example: *Kiedyś bardzo lubiłam wędliny z indyka*.
`I used to like cold cuts made from turkey.'

hyponym of {drób 1 - ‘poultry’}
hypernym of {indyczka 1 - ‘hen turkey’}

interlingual synonym of:
{**turkey 4**} (jedz) (semantic domain: ‘food’),
gloss: flesh of large domesticated fowl usually roasted
hyponym of {poultry 1}
hypernym of {hen turkey 1}

{drób 1} - I-syn {poultry 1}
{indyczka 1} - I-syn {hen turkey 1}

It seems that the link between the plWN synset {indyk 1} and the WN synset {turkey 4} yields a pair of very good translational equivalents. A subtle sense contrast can be tracked to the gloss of {turkey 4} which specifies that the meat is usually roasted, while no such information is included in the gloss of {indyk 1}. Yet it does not result in any substantial sense difference. The I-synonymy relation is established on the basis of correspondence in meaning which is mainly determined on the basis of correspondence in relation structure, (cf. Rudnicka et al. 2012). As demonstrated above, both hypernyms and hyponyms of the two synsets are also each other’s I-synonyms. Also, *PWN-Oxford Polish-English Dictionary* lists ‘turkey’ as an equivalent of *indyk* under a modifier *culin.* - ‘culinary’. It is the “first choice”, spontaneous equivalent that comes to mind of a bilingual speaker. Thus, the (two) lexical units: {indyk 1} and {turkey 4} also appear to function as cognitive equivalents.

In comparison to bi- and multilingual wordnet alignment among wordnets of other languages, the mapping between plWordNet and Princeton WordNet is unique with respect to a rich set of inter-wordnet interlingual relations applied. Apart from the simple equivalence relation that is I-synonymy, a set of more complex interlingual relations is employed (cf. Rudnicka et al. 2012). They are partly modelled on *Equivalence Relations* proposed within EuroWordNet project (cf. Vossen 2002), but they were not really used in any of EuroWordNet wordnets. The most frequent interlingual relation is I-hyponymy (with the reverse I-hypernymy relation). It is illustrated in example (5) for one Polish and two English synsets:

(5) {**palec 1**} (czc) ‘finger or toe; digit’ (semantic domain: ‘body parts’)
register: general; gloss: ‘u ludzi lub zwierząt’ - ‘of people or animals’
hyponym of {element anatomiczny 1 - ‘anatomical element’
hypernym of {duży palec 1 - ’big toe’}, {mały palec 1 - ‘little finger or toe’},
{palec wskazujący 1 - ‘index finger’}, {palec środkowy 1 - ‘middle finger’}, {serdeczny palec 1 - ‘ring finger’}

interlingual hypernym of:
{**finger 1**} (czc), (semantic domain: 'body parts')
gloss: any of the terminal members of the hand (sometimes excepting the thumb);
usage example: "her fingers were long and thin";
{**toe 1**} (czc), (semantic domain: 'body parts')
gloss: one of the digits of the foot;
{hammertoe 1} (czc) (semantic domain: 'body parts')

Polish, unlike English, does not lexicalise the semantic distinction between 'finger' and 'toe'. It has one word subsuming the meaning of the two English words[4]. When the context does not disambiguate its meaning, the Polish word can be post-modified with a prepositional phrase: *palec u ręki* - 'digit of a hand', or *palec u nogi* - 'digit of a foot'. Usually, it is not needed: *But mi obtarł palca*. - 'The shoe hurt my toe'. vs. *Włożył dwa palce do rękawiczki*. - 'He put two fingers into the glove'. Therefore, {finger 1} and {toe 1} are linked to {palec 1} via I-hyponymy relation. We ran a translation probability check in the Polish-English parallel corpus *Paralela*[5] (cf. Pęzik 2016) and we found the following results:

(6) *palec-finger* 1400
palec-toe 160
palec-hand 85
palec-thumb 40
palec-digit 0
palec-dactyl 0

Although an analysis of the sample of the results in (5) showed that quite a few of matches are false-positives (they appear in the parallel sentences, but the words are not necessarily translations of each other), there is clearly a tendency for *palec* to be translated as *finger*. This, in turn, is clearly related to the fact that, intuitively, we talk more frequently about fingers than about toes.

Another interlingual relation used in plWN-WN synset mapping is I-partial synonymy, devised to cater for cases of partial overlap of meanings and relation structures. It is exemplified below:

[4] Interestingly, Japanese also has one word for *finger* and *toe* - *yubi* 指.
[5] http://paralela.clarin-pl.eu/

(7) {**mebel 1** - ‘a piece of furniture’} (wytw) (semantic domain: ‘artefact’)
register: general;
gloss: *sprzęt użytkowy przeznaczony do wyposażenia wnętrz mieszkalnych i publicznych, posiada też walory dekoracyjne, reprezentacyjne; w odróżnieniu od stałych elementów wyposażenia wnętrz (schody, boazeria) meble są ruchomościami*
‘an instrumentality dedicated to furnishing private and public interiors, it also has decorative and representative value; in contrast to immovable elements of furnishing
(stairs, wainscot) furniture is movable’
usage example: *Kolejną kwestią sporną, dotyczącą mieszkanka, jest rozstaw mebli w kuchni.*
‘Another arguable issue concerning the little flat is the setup of furniture in the kitchen.’

hyponym of {element wyposażenia 1 - ‘an article of furnishing’}
hypernym of {biblioteczka 1 - ‘bookcase’},
{S mebel sypialniany 1 ‘a piece of bedroom furniture} (artificial synset) (and a few more)
meronym of {umeblowanie 2 - ‘furniture’}

interlingual partial-synonym of:
{**furniture 1**, piece of furniture 1, article of furniture 1} (semantic domain: ‘artefact’)
gloss: furnishings that make a room or other area ready for occupancy;
usage example: "they had too much furniture for the small apartment";
"there was only one piece of furniture in the room"

hyponym of {furnishing 1}
hypernym of {bookcase 1}, {bedroom furniture 1}, (and a few more)
interlingual hypernym of {umeblowanie 2 - ‘furniture’}

The Polish synset {mebel 1 - ‘a piece of furniture’} includes only one lexical unit in singular number and it is within singular tree (see its hypernym and hyponyms). Polish also has a mass term with the equivalent meaning placed in a separate synset - {umeblowanie 2 - ‘furniture’}, linked by a holonym relation to {mebel 1}. The most semantically close equivalent English synset subsumes both a mass term – *furniture* 1, together with two singular composite terms – *piece of furniture* 1 and *article of furniture* 1. It is located under a singular hypernym - {furnishing 1}, and has both singular and mass hyponyms - {bookcase 1} and {bedroom furniture 1}. After the analysis of the relation structures, glosses and usage examples of {mebel 1}, {umeblowanie 2} and {furniture 1, piece of furniture 1, article of furniture 1}, a lexicographer decided to link {mebel 1} via I-partial synonymy and {umeblowanie 2} via I-hyponymy to {furniture 1, piece of furniture 1, article of furniture 1}.

Clearly, stronger direct equivalence links could be established between the lexical units *mebel* 1 and *piece of furniture* 1, and between *umeblowanie* 2 and *furniture* 1.

3 Defining equivalence for sense mapping

So far we have seen that the current synset-level mapping between plWordNet and Princeton WordNet offers a lot of information with respect to potential translational equivalents between Polish and English, and yet, despite a rich system of inter-wordnet interlingual relations, it does not unequivocally signal the strength of the link of particular members of the mapped pair of synsets (see Example (2)). There are two main reasons behind such state of affairs. First, synsets differ in granularity: they may contain one or more lexical units which entails three options of matching: 1-to-1, 1-many, and many-to-many. Second, inter-wordnet interlingual relations are established largely on the basis of correspondence of relation structures between the two synsets. Hence, we deal with system correspondence rather than usage correspondence here (cf. Rudnicka et al 2017). This opens up the way to the possibility of occurrence of both cognitive and translational equivalents not only within I-synonymy, but also within I-hyponymy, I-hypernymy and I-partial synonymy (see Example (4)).

Therefore, in this paper we propose a strategy for signalling the strength of semantic link between lexical units (nouns) of plWordNet and Princeton WordNet. The starting point for creating an (additional), more fine-grained network of lexical unit links will be the existing network of interlingual relations between Polish and English synsets. These will be re-analysed from the perspective of cognitive and translational equivalence (described in Section 2). The candidate lexical units with a potential for stronger links will be identified and verified against a set of features that will be specified below.

In the first step, we will extract the following list of bilingual pairs from the mapped plWordNet:

(8) i. single lexical unit synsets linked by I-synonymy (*1-to-1* match)
ii. single and multiple lexical unit synsets linked by I-synonymy (*1-to-many* match)
iii. multiple lexical unit synsets linked by I-synonymy (*many-to-many* match)

Later, we will also extract similar lists for other interlingual relations such as I-hyponymy, I-hypernymy and I-partial synonymy. The lists will be then given to bilingual lexicographers. They will be asked to verify the strength and specificity of bilingual links between all member lexical units of the extracted pairs of synsets on the basis of detailed guidelines and they will introduce special new links where applicable.

The two types of strong equivalence described in Section 2, that is cognitive and translational, emphasise its two different aspects, namely meaning and use, respectively (cf. Piotrowski 2011; Adamska-Sałaciak 2014). For the purposes of (direct) lexical unit mapping, we would like to define two types of strong equivalence cross-cutting through these two aspects. The first type, *super-strong equivalence* would entail the agreement (identity and/or compatibility) in the possibly high number of features. The second type, *strong equivalence* would mean the agreement in the reasonably high member of features. For both, a common set of primary, always necessary to agree in, features will be specified. Below we present a proposal of *super-strong equivalence* set of features:

(9) **Super-strong** equivalence:

i. identity in **grammatical category** (given from the synset mapping)
ii. identity in **number**
iii. identity in **sense** (synset (and lexical unit) relation structure and gloss)
iv. identity in **register**
v. identity in **countability**
vi. compatibility in (semantic) **gender** (if relevant/applicable)
vii. **'first choice**' equivalent: listed first in bilingual dictionaries
viii. **bidirectional**
ix. **high translation probability** if it appears in a parallel corpus
x. **unique** for a single lexical unit

The set of super-strong equivalence features proposed in (8) includes: formal (morpho-syntactic) features such as grammatical category, number and countability; semantic features such as sense, register and (semantic) gender; and usage features such as dictionary listing, directionality of translation (substitution) and translation probability. The last feature is more of a requirement for lexicographers who will do the linking. It stipulates that there can be only one such relation per one lexical unit. Formal features are generally easy to determine. Besides, lexical unit mapping will draw on the results of synset mapping and interlingual synset relations obtain between synsets of the same grammatical category (except for cross-categorial synonymy, which will not be taken into account here). As for number and countability, plWordNet always puts singular and mass terms into separate synsets, but this does not always hold for Princeton WordNet (see Example (7) for a 'mixed' singular and mass synset).

Semantic features bring about more challenge. As for gender, both wordnets place feminine terms into separate synsets. Moreover, feminine gender is marked in plWordNet by the lexical unit relation *Żeńskość* - 'feminine gender'. Feminine terms are usually derived from masculine terms in Polish and it is a very productive process. If not derived, they are still located under a feminine gender hypernym. Apart from exceptional cases (e.g., *fiancé/fiancée, waiter/waitress*),

English does not mark morphological gender. To some extent, gender is lexicalised in English (e.g., *actor/actress*). Again, this a fairly productive derivation process. Feminine terms can be hyponyms of neutral/masculine terms in Princeton WordNet. Still, the information about feminine gender can be always found in the synset gloss.

On the other hand, the information about stylistic register is often absent from Princeton WordNet. Sometimes, stylistically marked lexical units appear in the same synsets with neutral, unmarked terms. This usually does not happen in plWN. Stylistically marked terms are put into separate synsets which are linked to their unmarked synonyms by the relation of *Bliskoznaczność* - 'Inter-register synonymy'. Moreover, the constructors of plWordNet have recently started to enrich lexical unit information with register (cf. the system of registers in plWordNet). Undoubtedly, the semantic description offered by wordnets is still to some extent limited and there are contrasts and gaps in lexical description between plWN and WN (also see below), so lexicographers will need to consult other lexical resources too.

Clearly, the key and most challenging feature to determine is identity in sense. The key sense denominator in a wordnet is a network of lexico-semantic relations. Both Princeton WordNet and plWordNet have a rich network of synset relations. plWordNet has also developed an equally rich network of lexical unit relations which are present, but much less numerous in the original WordNet. They add more specification to the semantic information of individual lexical units. Moreover, plWordNet provides glosses on the lexical unit level, although not all lexical units have been attributed with them by now. Originally, plWordNet did not include glosses at all. On the other hand, Princeton WordNet has glosses for all synsets. For some, there are also usage examples provided, similarly for plWordNet. Again, there has recently been a shift in plWordNet development process to move examples to lexical unit level. An important factor in determining the correspondence in relation structures will be the existing network of inter-wordnet interlingual synset relations between plWordNet and Princeton WordNet. It has to be taken into account in determining the meaning and degree of equivalence between a given pair of Polish and English lexical units, but, in general, lexical unit mapping needs to go beyond the existing synset mapping and show more detailed (subtle) sense and usage correspondences.

The last sub-set of features involves the potential of the use of equivalents in translation. It takes from the cognitive equivalent the "first choice" requirement. Since leaving it (purely) to the intuition of individual lexicographers would result in inconsistent results, we decided to treat bilingual dictionary listing as a kind of benchmark. It should be (preferably) listed first in bilingual dictionaries (for the specific sense of a headword). Moreover, a Polish-English pair of lexical units linked by super-strong equivalence should be bidirectional, that is, it should have

the same equivalent power in both translation directions (Polish-English/English-Polish). Again, this can be (at least partly) verified in bilingual dictionaries.

In addition, we decided to measure translation probability in a Polish-English parallel corpus: *Paralela* (Pęzik 2016) taking into consideration its limitations. The first one is the current size of the corpus, which includes 262 million words in 10,877,000 translation segments; as for the structure and composition of the corpus, legal texts (European Union legislation, proceedings of European Parliament etc.), press releases, medical texts (made available online by the European Medicine Agency) as well as film subtitles predominate therein (Pęzik 2016: 68). The second one is the level of alignment it offers, which is currently at the level of sentences (Pęzik 2016: 70), yet the search engine supports query-based word alignment and the equivalents of the source-language words (queries) are identified and ranked using the Dice coefficient (Pęzik 2016: 73). However, only a small part of the corpus was manually aligned (popular science texts from Academia and Center of Eastern Studies (OSW) as well as 114 literary novels), which means that there is still an area for improvement as regards an overall quality of alignment. In addition, there seem to be a few files where the text labelled Polish is in fact English or Russian, and a few places where the same text appears multiple times, so we will do some preprocessing to remove these, rejecting all sentence pairs not identified as (Polish, English) by langid.py (Lui and Baldwin 2011) and then removing any duplicate pairs.

For our purposes, we would ideally need a sense-level alignment (which would require prior bilingual word-sense disambiguation of corpus data). We will calculate ngram-based translation probabilities: the probability that contiguous chunk of text in one language is translated by a chunk in the other, using the anymalign tool (Lardilleux and Lepage, 2009), a lightweight aligner which calculates the probabilities based on random sampling. anymalign gives absolute frequency of the ngram pair as well as the probability of translation from source to target (e.g. what is the probability that an instance of *palec* being translated as *finge*r) and target to source (*finger* to *palec)*. A good pair should have high translation probabilities in both directions. Since we are interested in the lexical unit rather than the surface form, we will calculate the probabilities both for surface form of the words and the lemmatized version (where *fingers* is replaced by *finger*).

Note that pairs of equivalents extracted from parallel corpora are translational ones based on words: they tend to be context-specific and not necessarily identical in meaning. For our super-strong equivalence links, we really want items with identity in sense (as described above). Still, all this does not mean that data extracted from parallel corpora are of no use for us. It certainly shows translation/equivalent tendencies and professional lexicographers are usually able to disambiguate the meaning of a searched word on the basis of its context.

However, it is important to bear in mind that given the limitations of parallel corpora (in terms of structure, composition, representativeness, balance etc.), any tendencies in translation, revealed by, for example, translation probabilities, should be interpreted only with respect to texts found in a given parallel corpus rather than being extrapolated to the originals and their translations in their totality. In particular, a low translation probability does not mean that a pair is not a good equivalent, it may just be the case that this usage never appeared in the corpus.

Now, it may seem hard to find any pairs of Polish-English lexical units fulfilling all of the above described features. Our suggestions for super-strong equivalence instances go as follows: *indyk* 1 and *turkey* 4, *indyczka* 1 and *hen turkey* 1, *olej słonecznikowy* 1 and *sunflower oil*, *pies* 2 and *dog* 1, *pies domowy* 1 and *domestic dog* 1, *mebel* 1 and *piece of furniture* 1. A detailed description of these examples was presented in Section 2, examples (2), (3), (4) and (7).

The super-strong equivalence as defined above rests on the assumption that there is some commonality between lexical structures of the two languages. It assumes that some concepts are lexicalised to a very similar degree. However, it is generally accepted that no universal lexical structure exists (cf. von Fintel and Matthewson 2008) and there are a number of lexical gaps and mismatches between languages (cf. Svensen 2009). This was confirmed in the process of synset mapping between plWordNet and Princeton WordNet (cf. Rudnicka et al. 2016). The most frequent interlingual relation is I-hyponymy. Still, despite these differences, translations are made, dictionaries are constructed and speakers of different languages do communicate. To capture meaning and usage correspondences that are still very strong, yet not complete in all aspects, we would like to propose the *strong equivalence* relation. It will share the features with super-strong equivalence, yet the requirements for some of them will be slightly relaxed in comparison with the former, see below:

(10) **Strong equivalence** features:

i. identity in **grammatical category** (given from the synset mapping)
ii. identity in **number**
iii. largely compatible in **sense** (synset (and lexical unit) relation structure and gloss)
iv. identity in **register**
v. identity in **countability**
vi. compatibility in (semantic) **gender** (if relevant/applicable)
vii. often ‘first choice’ equivalent: often listed first in bilingual dictionaries
viii. can be either uni or bidirectional
ix. preferably high translation probability in a parallel corpus
x. need not be unique for a single lexical unit

The strong equivalence specification differs from the super-strong equivalence in a number of aspects. The values for formal features are alike. In terms of semantics, it allows for large compatibility in sense instead of identity. In terms of usage, unidirectionality is accepted, along with bidirectionality, and the 'first choice' is a preference, not an obligation. Consequently, uniqueness of a link also need not to be observed. Strong equivalence examples are as follows: *palec* 1 and *finger* 1, *palec* 1 and *toe* 1, *olej słonecznikowy* 1 and *sunflower-seed oil* 1, *mebel* 1 and *article of furniture* 1.

Hence, the proposed lexical unit mapping will be partly based/modelled on the existing system of interlingual relations between synsets. It is aimed to enrich the current system of bilingual links, and not to repeat the same information. Therefore, we will introduce new, more detailed links only in the cases where there is enough motivation for such a decision. The remaining, not directly linked lexical units from synsets otherwise linked by I-synonymy, I-partial synonymy, I-hyponymy and I-hypernymy, will be argued to have *implied weak translational* equivalence. That basically means that they are *potential* translational equivalents depending on context.

4 Conclusions

In this paper, we have proposed a strategy of linking plWordNet and Princeton WordNet at the level of its smallest building parts - lexical units. The proposed sense-level matching may be viewed as a kind of extension of the existing inter-lingual mapping between synsets. Capitalizing on equivalence types from (traditional) lexicographic literature, we have defined two types of links between lexical units: *super-strong equivalence* and *strong equivalence*. They are aimed to cross-cut through the import of the so-called cognitive and translational equivalents. They share a set of formal, semantic and usage features. The requirements for strong equivalence are only partly loosened in comparison to super-strong one. The sense-level mapping will draw on the results of synset-level mapping to the extent that lists of pairs of synsets linked by I-synonymy, I-partial synonymy, I-hyponymy and I-hypernymy will be extracted and further grouped into single lexical unit pairs (1-1), multiple lexical unit pairs (many-many) and single to multiple lexical unit pair (1-many). The lists and sets of equivalence features will be next given to trained bilingual lexicographers who will verify the strength of correspondence between bilingual pairs of lexical units (of each synset pair) and will introduce new super-strong and strong equivalence links where applicable. They will be encouraged to consult various lexical resources as well as check translation probabilities in parallel corpora. The remaining lexical units of the earlier mapped synset pair are argued to have the *implied weak translational*

link which signals a potential for becoming a translational equivalent in an appropriate context.

The proposed strategy paves the way for creating a more fine-grained, detailed network of interlingual links of a great potential of application in both natural language processing tasks (e.g. automatic translation, bilingual word-sense disambiguation, sense-level alignment of parallel corpora), manual translation and language learning, among others.

Acknowledgements

The paper is the result of works carried out within the project funded by the National Science Centre (*Narodowe Centrum Nauki*), Poland, under the grant agreement no: UMO-2015/18/M/HS2/00100.

References

Adamska-Sałaciak, Arleta. 2010. Examining equivalence. *International Journal of Lexicography* 23(4). 387–409.

Adamska-Sałaciak, Arleta. 2013. Issues in compiling bilingual dictionaries. In Howard Jackson (ed.), *The Bloomsbury companion to lexicography*, 213–231. London: Bloomsbury.

Adamska-Sałaciak, Arleta. 2014. Bilingual lexicography: translation dictionaries. In Patrick Hanks & Gilles-Maurice de Schryver (eds.), *International handbook of modern lexis and lexicography*, 1–11. Springer-Verlag: Berlin-Heidelberg.

Bentivogli, Luisa & Emanuele Pianta. 2004. Extending WordNet with Syntagmatic Information. In *Proceedings of the Second Global WordNet Conference*, Brno, Czech Republic, January 20–23, 2004, 47–53.

Crenn, Tiphaine. 1996. *Register and register labelling in dictionaries*. Ottawa: University of Ottawa.

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

von Fintel, Kai & Lisa Matthewson. 2008. Universals in Semantics. *The Linguistic Review* 25(1-2). 139–201.

Fišer, Darja & Benoıt Sagot. 2015. Constructing a poor man's wordnet in a resource-rich world. *Language Resources & Evaluation* 49(3). 601–635.

Hamp, Birgit & Helmut Feldweg. 1997. *GermaNet – a Lexical Semantic Net for German.* In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo & Yorick Wilks (eds.), *Proceedings of ACL workshop Automatic*

Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 9–15. Madrid: ACL.

Héja, Enikő. 2016. Revisiting translational equivalence: contributions from data-driven bilingual lexicography. *International Journal of Lexicography*, ecw032.

Kamiński, Mariusz. 2016. Towards successful communication between the dictionary and the user. In Anna Kuzio, Jolanta Kowal & Mirosława Wawrzak-Chodaczek (eds.), *Social communication in the real and virtual world*. Vol. 1., 73–91. Saarbrücken: LAP LAMBERT Academic Publishing.

Lardilleux, Adrien & Yves Lepage. 2009. Sampling-based multilingual alignment. *International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria, 214–218. Retrieved from: https://hal.archives-ouvertes.fr/hal-00439789/document.

Lew, Robert. 2013. Identifying, ordering and defining senses. In Howard Jackson (ed.), *The Bloomsbury companion to lexicography*, 284–302. London: Bloomsbury.

Lindén, Krister & Lauri Carlson. 2010. FinnWordNet – WordNet påfinska via översättning, *LexicoNordica – Nordic Journal of Lexicography*, 17. 119–140. [English translation 'FinnWordNet – Finnish Word-Net by translation']. Retrieved from: http://www.ling.helsinki.fi/~klinden/pubs/FinnWordnetInLexicoNordica-en.pdf.

Lui, Marco & Timothy Baldwin. 2011. Cross-domain Feature Selection for Language Identification, In *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011*), Chiang Mai, Thailand. 553–561. Retrieved from: http://www.aclweb.org/anthology/I11-1062.

Maziarz, Marek, Maciej Piasecki & Stanisław Szpakowicz. 2013a. The chicken-and-egg problem in wordnet design: synonymy, synsets and constitutive relations. *Language Resources and Evaluation* 47(3). 769–796.

Maziarz, Marek, Maciej Piasecki & Stanisław Szpakowicz. 2015. The System of Register Labels in plWordNet. *Cognitive Studies* 15. 161–175.

Pęzik, Piotr. 2016. Exploring phraseological equivalence with Paralela. In Ewa Gruszczyńska & Agnieszka Leńko-Szymańska (eds.), *Polish-Language Parallel Corpora*, 67–81. Warszawa: Instytut Lingwistyki Stosowanej UW.

Piasecki, Maciej, Stanisław Szpakowicz & Bartosz Broda 2009. *A wordnet from the ground up.* Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Piasecki, Maciej, Marek Maziarz, Ewa Rudnicka, Agnieszka Dziob & Paweł Kędzia. 2017, in print. plWordnet – a Large Corpus-Based Wordnet of Polish. *Linguistic Issues in Language Technology*.

Piotrowski, Tadeusz. 2011a. Ekwiwalencja w słownikach dwujęzycznych. In Wojciech Chlebda (ed.), *Na tropach translatów: w poszukiwaniu odpowiedników przekładowych*, 45–70. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Piotrowski, Tadeusz. 2011b. Tertium comparationis w przekładoznawstwie. In Piotr Stalmaszczyk (ed.), *Metodologie językoznawstwa. Od ontologii do pragmatyki*, 175–192. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Rudnicka, Ewa, Marek Maziarz, Maciej Piasecki & Stanisław Szpakowicz 2012. A strategy of mapping Polish WordNet onto Princeton WordNet. In *Proceedings of COLING 2012*. Retrieved from: www.aclweb.org/anthology/C12-2101.

Rudnicka, Ewa, Wojciech Witkowski & Michał Kaliński. 2015. a semi-automatic adjective mapping between plWordNet and Princeton WordNet. In: Pavel Kral & Vaclav Matousek (eds.), *Text, speech, dialogue*, 360–368. Berlin: Springer.

Rudnicka, Ewa, Wojciech Witkowski & Łukasz Grabowski. 2016. Towards a methodology for filtering out gaps and mismatches across wordnets: the case of noun synsets in plWordNet and Princeton WordNet. In Verginica Barbu Mititelu, Corina Forascu, Christiane Fellbaum & Piek Vossen (eds.), *Proceedings of the Eighth International Global WordNet Conference 2016*, 27–30 Jan 2016, Bucharest, Romania, 344–351. Retrieved from: http://gwc2016.racai.ro/procedings.pdf

Rudnicka, Ewa, Maciej Piasecki, Tadeusz Piotrowski, Łukasz Grabowski & Francis Bond. 2017, in print. Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies* 17.

Rudnicka, Ewa, Maciej Piasecki & Wojciech Witkowski. 2017, in print. enWordnet – a mapping-based extension of Princeton WordNet. *Linguistic Issues in Language Technology.*

Svensen, Bo. 2009. *A Handbook of lexicography. The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Taylor, John. 2012. *The mental corpus. How language is represented in the mind*. Oxford: Oxford University Press.

Vossen, Piek (ed.). 2002. *EuroWordNet general documentation*, Version 3. Retrieved from: http://www.vossen.info/docs/2002/EWNGeneral.pdf

Yong, Heming & Jing Peng. 2007. *Bilingual lexicography from a communicative perspective*. Amsterdam: John Benjamins.

About the Authors

Ewa Rudnicka is a Research Associate at the Department of Computer Science and Management, Wroclaw University of Technology, Poland. Her research interests include computational bilingual lexicography, comparative linguistics, formal semantics, translation studies. She is the coordinator of the process of mapping plWordNet onto Princeton WordNet. She is a member of G4.19. Language Technology and Computational Linguistic Research Group.

Francis Bond is an Associate Professor at the Division of Linguistics and Multilingual Studies, Nanyang Technological University, Singapore. He worked on machine translation and natural language understanding in Japan, first at Nippon Telegraph and Telephone Corporation and then at the National Institute of Information and Communications Technology, where his focus was on open source natural language processing. He is an active member of the Deep Linguistic Processing with HPSG Initiative (DELPH-IN) and the Global WordNet Association. His main research interest is in natural language understanding. Francis has developed and released wordnets for Chinese, Japanese, Malay and Indonesian and coordinates the Open Multilingual Wordnet.

Łukasz Grabowski is an Associate Professor at the Institute of English, University of Opole, Poland. His research interests include corpus linguistics, phraseology, formulaic language, translation studies and lexicography. He is also interested in computer-assisted methods of text analysis. He has published internationally in *International Journal of Corpus Linguistics* and *English for Specific Purposes*, among others. He is also Managing Editor of the journal *Explorations: A Journal of Language and Literature*.

Maciej Piasecki is an Associate Professor at the Department of Computer Science and Management, Wroclaw University of Technology, Poland. He is a Polish National Coordinator of CLARIN ERIC (www.clarin.eu) and a member of Global WordNet Association Board. He has been an initiator and is the leader of plWordNet project (a large wordnet of Polish) and is the leader of G 4.19. Language Technology and Computational Linguistic Research Group. His research interests cover different areas of natural language processing and engineering, computational lexicography, data extraction and information retrieval.

Tadeusz Piotrowski is a Professor at the English Department, University of Wrocław, Poland. His research interests include theory, practice, and history of monolingual and bilingual lexicography and dictionaries, corpus linguistics, translation studies, participated in most major bilingual dictionary projects in Poland, working with such companies as PWN, OUP, Pons-Klett, Langenscheidt, Prószyński, Wiedza Powszechna, Kościuszko Foundation, and wrote a number of dictionaries for Spotkania. He is also interested in computational lexicography and computer-assisted text analysis. He published three books and about 200 papers.

Address
Ewa Rudnicka
Department of Computer Science and Management
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland

e-mail: ewa.rudnicka@pwr.edu.pl