Lexical Perspective on Wordnet to Wordnet Mapping

Ewa Rudnicka, [♡] ***Francis Bond,*** [♠]
Łukasz Grabowski, [♣] ***Maciej Piasecki*** [♡] ***and Tadeusz Piotrowski*** [♢]
[♡]Wrocław University of Technology
[♠]Nanyang Technological University, Singapore
[♣] University of Opole
[♢]University of Wrocław
{ewa.rudnicka,maciej.piasecki}@pwr.edu.pl, bond@ieee.org, lukasz@uni.opole.pl, tadeusz.piotrowski@uwr.edu.pl

Abstract

The paper presents a feature-based model of equivalence targeted at (manual) sense linking between Princeton WordNet and plWordNet. The model incorporates insights from lexicographic and translation theories on bilingual equivalence and draws on the results of earlier synset-level mapping of nouns between Princeton WordNet and plWordNet. It takes into account all basic aspects of language such as form, meaning and function and supplements them with (parallel) corpus frequency and translatability. Three types of equivalence are distinguished, namely strong, regular and weak depending on the conformity with the proposed features. The presented solutions are language-neutral and they can be easily applied to language pairs other than Polish and English. Sense-level mapping is a more fine-grained mapping than the existing synset mappings and is thus of great potential to human and machine translation.

1 Introduction

Currently, bi- and multilingual wordnets are most commonly inter-linked on the synset level, (e.g., Bond and Foster, 2013). Synsets can be composed of one or more lexical units (lemma-PoS-synset triples, also called senses; henceforth, LUs), so such inter-wordnet links may be of three types: *1-to-1* sense link (between two synsets each built of a single LU); *1-to-many* sense link (between two synsets, one built of a single LU, the other of more than one); and *many-to-many* sense link (between two multiple-LU synsets). The (large) majority of inter-linked wordnets use one simple equivalence relation to connect their synsets (effectively synonymy). If, due to substantial differences between languages, such a link cannot be introduced, sometimes *artificial* synsets are created to provide equivalents (e.g., Bentivogli and Pianta, 2004; Lindén and Carlson, 2010). When we consider 1-to-many and many-to-many sense links, the question arises whether the correspondence between all their component LUs is of the same strength. Basic principles of language economy state that within one language there should not exist two different forms that share identical function and meaning, so there have to be slight differences between component LUs of a given synset, and even larger differences between the LUs from two synsets representing two different languages (even if those synsets are linked by I-synonymy). Existing research on inter-wordnet mapping between plWordNet (Maziarz et al., 2016) and Princeton WordNet (Fellbaum, 1998), especially 1-to-many and many-to many sense links, has shown the potential for creating stronger links between some LUs from a given pair of synsets (Rudnicka et al., 2016). To give an example, in the pair of synsets: {***złoto***$_{n:3}$, ***Au***$_{n:1}$}^{PL}I-syn {***gold***$_{n:3}$, ***Au***$_{n:1}$, ***atomic number 79***$_{n:1}$}^{EN}— ***złoto***$_{n:3}{}^{PL}$and ***gold***$_{n:3}{}^{EN}$and ***Au***$_{n:1}{}^{PL}$and ***Au***$_{n:1}{}^{EN}$seem the best-fitted equivalents due to the agreement not only in sense, but also in register. The words from the first pair belong to the general register, while the ones from the second pair are from the specialist register. Bi- and multilingual wordnets are used by translators who would certainly appreciate such a more detailed mapping.

2 Background

Equivalence is a popular concept used in, among others, translation studies and bilingual lexicography – see Rudnicka et al. (2017b) for a more detailed discussion, also regarding typologies of

equivalence). The concept has many faces depending on which features of language or texts researchers focus on. For example, one may find binary oppositions such as, for instance, natural and directional equivalence, semantic and pragmatic equivalence, or full and partial equivalence, e.g. Pym (2007); Svensen (2009). When studying recent approaches to equivalence developed in the field of bilingual lexicography, one may also find a distinction between cognitive and translational equivalence (Adamska-Sałaciak, 2010; Heja, 2016). Cognitive equivalents are typically general ones; they first come to the mind of a language user (even without any context) and when it comes to translation they may fit many contexts. Translational equivalents, which may be extracted from corpus data, may be less obvious and sometimes they may slightly differ in their basic meaning; however, they may fit more specific contexts. In Rudnicka et al. (2017a), we analysed basic equivalence types from translation and lexicographic literature and verified their relevance for synset-level wordnet mapping. We assumed that LUs in the linked plWN-PWN synsets can be treated as bilingual dictionary data. We checked if pairs of LUs might be treated as cognitive and translational equivalents depending on their frequency of use as equivalents in translation in a particular co-text and context. We put forward an initial proposal of sense-level mapping designed to cross-cut through cognitive and translational equivalence. In this paper we present an extended and verified version of our initial proposal with carefully defined equivalence features, equivalence types and a sense-level linking procedure supported by a number of examples. At this point, it is important to note that the term equivalence has been also used in the context of wordnets; more precisely, it was first used in the wordnet world to name a set of inter-lingual relations holding between synsets in the EuroWordNet project (Vossen, 2002, p.:38). Inter-lingual synonymy was defined as a simple equivalence relation "which only holds if there is 1-to-1 mapping between synsets". The remaining types of inter-lingual relations were called Complex Equivalence relations and allowed to obtain between one-to-many and many-to-many synset pairs. Since many of EuroWordNet wordnets relied on translation approaches, many of the senses can be translational equivalents. In designing a strategy for mapping synsets between Princeton WordNet and plWordNet, Rudnicka et al. (2012) built on this proposal in the set of I-relations.

Currently, the overall size of plWordNet amounts to 217,426 synsets, 282,749 senses (lexical units) and 190,555 lemmas and these numbers are constantly growing. The synset mapping between plWordNet and Princeton WordNet encompasses 230,185 links, with 43,740 inter-lingual synonymy links. The number of Polish synsets with at least one inter-lingual relation is 177,634 (only I-synonymy is unique to a synset pair). The majority of I-links form noun links, 122,811 instances covering about 92% of Polish noun synsets (132,380 in total), the next are adjective links: 45,282 instances, covering 96% (46,721 in total), and last come adverb links with 9,541 instances, covering 84% of Polish adverb synsets (11,256 in total). At the present stage there are no inter-lingual links between verbal synsets (27,069), but we are working on the mapping procedure for them. Looking from the Princeton WordNet direction, we have mapped 80% of noun synsets (72,621), 43% of adjective synsets (7,905), and 47% of adverb synsets (1,737).

Since nouns are the most stable semantic category, we have decided to make them the starting point for our procedure for sense mapping. Other categories may require category- specific treatment which is outside the scope of the present paper.

3 Equivalence Features

In this section we discuss a set of features that will determine the strength of equivalence holding between (particular) LUs from the mapped Polish and English synsets. Each feature will be followed by a short definition and examples. First, we will look at *formal features*, such as grammatical category, number, countability and gender. Next, we will delve into *semantic* and *pragmatic* ones, such as sense, lexicalisation (of concepts), register, collocations, co-text and context. Finally, we will consider translatability based on dictionary listing and translation equivalences extracted from the Polish-English parallel corpus *Paralela*[1] (Pęzik, 2016).

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

3.1 Formal features

The first, basic, formal feature is identity in *grammatical category* between source and target LUs. Since sense-level mapping will be based on the results of an earlier synset-level mapping, this feature will be treated as 'given'. The interlingual relations that will be taken into consideration include I-synonymy, I-partial synonymy, I-hyponymy and I-hypernymy, all of which hold between the same part of speech synsets. Our focus will be the relations between nouns.

The more interesting formal features are *number* and *countability*. For regular, countable nouns, agreement in these features is usually also given, because both in plWordNet and in Princeton WordNet lemmas appear in singular form. Still, some cases of 'mixed' Princeton WordNet-synsets were already tracked e.g. {***dumpling***$_{n:1}$, ***dumplings***$_{n:1}$} (Rudnicka et al., 2012, 2016). Such mixed synsets currently serve as inter-lingual hypernyms for both singular and plural Polish synsets e.g.{***pieróg***$_{n:2}$, ***pierog***$_{n:1}$} 'dumpling' or {***pierogi ruskie***$_{n:1}$} 'Russian dumplings'. Still, sense level mapping will allow to resolve such inconsistencies in the synset built-up. In regular cases, the agreement in number will always be observed in the mapping.

A different case are pluralia and singularia tantum that have regular countable nouns as equivalents in another language such as, for instance, {***drzwi***$_{n:1}$} 'door^{pl}' I-syn {***door***$_{n:1}$}, {***grabie***$_{n:1}$} 'rake^{pl}' I-syn {***rake***$_{n:1}$}, {***centrala***$_{n:2}$} I-syn {***headquarters***$_{n:1}$}, or {***stajnia Augiasza***$_{n:1}$} 'Augeas' stable' I-syn {***Augean stables***$_{n:1}$}. A re-analysis of their relation structures and glosses shows very close meaning correspondence and leads to the conclusion that the difference in number is only a difference in grammaticalisation of the same concept. A similar case are regular nouns mapped to mass or group nouns, such as {***grzmot***$_{n:1}$} 'thunder^{sg}' I-syn {***thunder***$_{n:2}$} or {***błyskawica***$_{n:1}$} 'lightning^{sg}' I-partial-syn {***lightning***$_{n:2}$}. There are also cases of pluralia tantum mapped to uncountable nouns e.g. {***wagary***$_{n:1}$} 'truancy^{pl}' I-syn {***truancy***$_{n:1}$, ***hooky***$_{n:1}$}. On the basis of the above examples, we want to argue that identity in number and countability is an important criterion only in the case of regular, countable nouns. Cases of singularia and pluralia tantum should be dealt with on an individual basis. The features may gain more importance in the case of 1-to-many and many-to-many sense pairs e.g. {***odwiedziny***$_{n:1}$, ***wizyta***$_{n:1}$} I-syn {***visit***$_{n:1}$}.

The last formal feature is *gender*. One of the typical differences between a morphologically synthetic language (Polish) and an analytical one (English) is the degree of gender lexicalisation. Gender is systematically lexicalised in Polish, marked by derivational suffixes e.g. *nauczyciel* 'teacher' and *nauczycielka* 'female teacher', while it is much less lexicalised in English — it is sometimes signalled by derivational suffixes e.g. *emperor – empress*, sometimes by different, derivationally unrelated words e.g. *mare – stallion*. We suggest to constrain sense links with gender identity between LUs only in cases where both languages lexicalise the distinction, while in the remaining, contrasting cases mark the equivalence as slightly weaker than in former ones. Such a proposal is motivated by the fact that we consider information about natural gender to be an additional meaning component. Thus, we get very close correspondence between LUs in the following synset pairs: {***ogier***$_{n:1}$} I-syn {***stallion***$_{n:1}$, ***entire***$_{n:1}$} and {***klacz***$_{n:1}$} I-syn {***mare***$_{n:1}$, ***female horse***$_{n:1}$}, while just close correspondence between the pairs {***nauczyciel***$_{n:1}$} and {***nauczycielka***$_{n:1}$} I-hypo to {***teacher***$_{n:1}$}.

3.2 Semantic features

As already alluded in the previous section, the key denominator for LU mapping will be the correspondence in sense. By definition, the component LUs of a given synset do share the same (basic) meaning (Fellbaum, 1998). Still, in such a model, some more subtle meaning distinctions may not be captured, such as shades of meaning going beyond Leibniz's (1704) truth-conditional understanding of synonymy. Other factors that determine meaning are similarities and differences in lexicalisation of concepts, register, style, typical co-texts and contexts. They need not be of importance in some language processing tasks, but are always important for a translator. Therefore, the proposed sense-level mapping aims to go beyond the existing synset level mapping in the granularity and specificity of links. Currently, the I-synonymy link between synsets signals their correspondence in sense based mainly on their synset relation network (and partly on glosses and examples of use that come with synsets in Princeton WordNet and with LUs in plWordNet). In LU

mapping, we would like to re-analyse the existing inter-lingual synset links, and wherever possible, establish sense links of a stronger character. We see the potential for stronger sense links especially in the case of 1-to-many and many-to-many sense pairs. For these purposes, we will need to consult external resources such as mono- and bilingual dictionaries, encyclopaedia, and mono- and parallel corpora.

An example of 1-to-many sense pair is the Polish synset {***narzeczona***$_{n:1}$} 'fiancee' linked via I-synonymy to the English synset {***fiancee***$_{n:1}$, ***bride-to-be***$_{n:1}$}. The Polish gloss can be translated as "a woman who obliged herself to marry a concrete man (her fiance), made him such a promise", while the English one is just "a woman who is engaged to be married". Having consulted a couple of monolingual English dictionaries Cobuild (2012); CALD (2013); LDCE (2014), we find that *fiancée* is defined as "the woman that a man is engaged to/going to marry", while *bride-to-be* as "a woman who is going to be married soon". Clearly, there is an additional meaning component in the case of *bride-to-be*, namely *soon*, not included in the general synset gloss. The synset gloss corresponds more closely to the dictionary definitions of *fiancée* and to the Polish gloss of ***narzeczona***$_{n:1}$. Therefore, there is a stronger link between lexical units ***narzeczona***$_{n:1}$ and ***fiancée***$_{n:1}$ than between ***narzeczona***$_{n:1}$ and ***bride-to-be***$_{n:1}$.

An important factor influencing equivalence between LUs of the two languages are similarities and differences in lexicalisation of the same concepts. These will be judged by comparing the denotations of bilingual pairs of LUs. An example is the Polish word ***zabytek***$_{n:2}$ 'historic monument' with the gloss: "stary budynek, przedmiot" 'an old building, artefact' which denotes anything of historic value no matter of its size. There is no direct equivalent of this word in English. One has to use a different noun depending on the size of an object e.g. *historic monument*, *building*, *site*, *landmark*. The Princeton WordNet synset with the closest meaning is {***monument***$_{n:2}$} with the following gloss: "an important site that is marked and preserved as public property", an instance hyponym {***Stonenhenge***$_{pn:1}$} and a hyponym {***market cross***$_{n:1}$}. The two synsets {***zabytek***$_{n:2}$} and {***monument***$_{n:2}$} are linked by I-partial synonymy. In some contexts ***monument***$_{n:2}$ will be the best translation of ***zabytek***$_{n:2}$, yet their overall meaning correspondence is partial.

Another area to look for more meaning specification is *register*. More precisely, registers are marked only for very few Princeton WordNet synsets by means of the *Domain Usage* relation, of which a couple of specifiers are of interest to us, namely {***archaism***$_{n:1}$}, {***colloquialism***$_{n:1}$}, {***disparagement***$_{n:1}$}, {***ethnic slur***$_{n:1}$}, {***formality***$_{n:3}$}, {***vulgarism***$_{n:1}$} and {***slang***$_{n:2}$}. In plWordNet registers are marked for lexical units and the following ones are distinguished: *general*, *official*, *specialist*, *literary*, *colloquial*, *common*, *vulgar*, *obsolete*, *regional*, *slang/argot* and *non-normative*. There are some cases of correspondence in register systems between English and Polish e.g. {***big fish***$_{n:1}$, ...} linked by Domain Usage relation to {***colloquialism***$_{n:1}$} and via I-partial synonymy relation to the Polish synset {***gruba ryba***$_{n:1}$ 'big fish', ***ważniak***$_{n:2}$ 'VIP'} with both its LUs marked for the colloquial register. However, such simple cases are rare. Both in Princeton WordNet and in plWordNet, LUs of different registers can co-occur in the same synset. However, in the latter only LUs of *compatible* register can be grouped in one synset or linked by some relation, e.g. hypernymy. A set of rules was defined for this purpose in plWordNet (Maziarz et al., 2014), while this aspect is largely unconstrained in Princeton WordNet. General, specialist, literary, and official registers can co-occur in one synset; the same holds for general and colloquial ones (provided that that specialist, literary and official are not found in the same synset). Colloquial, common and vulgar can also come together. On the other hand, regional, obsolete, slang/argot and non-normative always come on their own. An example is the Polish synset {***okulary***$_{n:1}$ 'glasses': *general* register, ***patrzałki***$_{n:1}$, ***szkła***$_{n:1}$ 'specs': *colloquial* register, ***binokle***$_{n:2}$ 'eyeglasses': *colloquial* register}. ***okulary***$_{n:1}$'s gloss is translated to ""an optical device built of a pair of lens and a frame enabling fitting the lens in front of the eyes most often by ear arms, usually used to correct sight acuity, weakened by an illness, injury or age)'.. It is linked by I-synonymy relation to the English synset {***spectacles***$_{n:1}$, ***specs***$_{n:1}$, ***eyeglasses***$_{n:1}$, ***glasses***$_{n:1}$} "(plural) optical instrument consisting of a frame that holds a pair of lenses for correcting defective vision". There is no information about register for the Princeton WordNet synset. Still, when we look

up its component LUs in English dictionaries we find that *spectacles* is classified as either formal or old-fashioned, *specs* as informal and *eyeglasses* as North American. That suggests a strong link between ***okulary***$_{n:1}$ with ***glasses***$_{n:1}$ (both of a general register), and possibly also with ***eyeglasses***$_{n:1}$ (though maybe by a slightly weaker link), while ***patrzałki***$_{n:1}$ and ***szkła***$_{n:1}$ with ***specs***$_{n:1}$ (all of an informal or colloquial register). In fact, the Polish word ***binokle***$_{n:2}$ marked with a colloquial register also has an old-fashioned flavour, which makes it a good equivalent for the English ***spectacles***$_{n:1}$.

An important means for disambiguating sense are *collocations*, *co-text* (co-occurring words and text fragments) and *context* (type of situation, speaker, target audience, purpose of communication, style etc.). Words with the same meaning that appear in similar language environments in two languages tend to be equivalents of each other. It can be illustrated by LUs from the following pair of synsets: {***centrala***$_{n:2}$} linked via I-synonymy to {***headquarters***$_{n:1}$, ***office***$_{n:1}$, ***main office***$_{n:1}$, ***home office***$_{n:2}$, ***home base***$_{n:2}$}. The pair of LUs ***centrala***$_{n:2}$ – literary a noun LU derived from the adjective *centralny* 'central'– and ***headquarters***$_{n:1}$ gets 40 hits in the *Paralela* corpus and a couple of concordances illustrating the use of these two equivalents in their co-text can be distinguished, e.g.:

- *Jesienią 2007 r. duńska centrala firmy Arriva poszukiwała ponad 400 kierowców autobusów...*
  'In the autumn of 2007, Arriva's Danish headquarters were looking for more than 400 bus drivers...'

- *Ponieważ jej europejska centrala znajduje się w Irlandii, ...*
  'As their European headquarters is located in Ireland, ...'

- *Do pierwszego sprawozdania centrala wydała krótki komentarz,...* 'Headquarters commented briefly on the first report, ...'

Other LUs in the English synset (***central office***$_{n:1}$, ***main office***$_{n:1}$, ***home office***$_{n:2}$, and ***home base***$_{n:2}$) either do not appear in a pair with *centrala* or are quite rare.

3.3 Translatability

We have already seen in the previous section that dictionaries and corpora are indispensable resources in determining many features of equivalence, because they provide different types of information that may be missing in wordnets (e.g. register, collocations or typical co-text or contexts). In the process of construction of contemporary bilingual dictionaries a lot of emphasis is put on the translatability of the provided equivalents (e.g., Zgusta, 1971), with better translation equivalences listed first. Therefore, we would like to suggest that *dictionary listing* be treated as one of the indicators of the strength of equivalence between LUs. The main Polish-English/English-Polish dictionaries to be consulted will be PWN-Oxford (2007), Collins-YDP (1997) and Słownik-Kościuszkowski (2014). An issue that immediately emerges here is directionality of translation. It is known that not all equivalents work equally well both ways, that is from L1 to L2 and from L2 to L1. It can be verified by the so-called back-translation, also using dictionaries. In the extreme case it there is not always an equivalent provided for a headword when you try to back translate.

Translation theorists distinguish between natural equivalence and directional equivalence. According to Pym (2007), natural equivalence describes the correspondence between words, expressions or text chunks on all dimensions of meaning. It typically concerns terminology (e.g. {***duck***$_{n:1}$} I-syn {***kaczka***$_{n:1}$} 'duck', both belonging to the semantic domain *animal*), prefabricated chunks of texts and specialized uses of words (e.g. *whereas – zważywszy, że* as found in certain legal texts), so it seems to exist prior to translation. On the other hand, directional equivalence refers to situations when translators actively search for equivalents of source words in the target language (often in cases of lexical or cultural gaps), so it is by definition uni-directional or one-way. An example is the Polish synset {***stachanowiec***$_{n:1}$, ***przodownik pracy***$_{n:1}$} whose gloss translates to 'in the Eastern Block countries: a person competing for a title of a most efficient worker". It is linked via I-hyponymy to the English synset {***toiler***$_{n:1}$} gloss: "one who works strenuously". As shown by the gloss, *stachanowiec* is a typical cultural gap; the concept is specific to Eastern Block countries. Its I-hypernym, *toiler* can serve as a translational equivalent from Polish to English, yet back-translation does not work in this case (cf Techland-Dictionary (2006): *toiler – człowiek ciężkiej pracy* 'a man of hard work'.)

4 Equivalence types

Relying on equivalence features described in the previous section, we will define three equivalence types of a variable strength: *strong*, *regular* and *weak* (implied). The categorisation to a given type will be based on values of features a bilingual pair of LUs will agree in. The types will be later reflected in three kinds of links between LUs.

Some features will be agreed across all types, while some other feature will differ. Summing up the discussion in Section 3.1, there will always be an agreement in grammatical category (only noun-to noun pairs are taken into consideration) and in most cases in number, countability and gender. Instances of pluralia and singularia tantum as well as count-to-mass mappings will be dealt on an individual basis – the agreement will not always have to hold. Cases of lexicalised natural gender in Polish will be treated in a similar way.

4.1 Strong equivalence

By its very name, the strong equivalence will be the strongest type of link. It will require identity in sense, similarity in lexicalisation of concepts, compatibility in register, a shared set of typical co-texts, dictionary listing (preferably as the first equivalent), bidirectionality (but not uniqueness) of translation and, preferably, frequent parallel corpora hits. The most suitable candidates for such strong correspondence are LUs from one element (LU) synsets linked via I-synonymy synset relation. A couple of examples are given below (for their full descriptions see Sections 3.1 and 3.2):

- ***drzwi***$_{n:1}$ I-syn ***door***$_{n:1}$
- ***grzmot***$_{n:1}$ I-syn ***thunder***$_{n:2}$

  All strong because of identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

The second group of examples to consider are one-to-many sense pairs of synsets linked via I-synonymy. It is likely that there will be at least one pair of LUs that will meet the strong equivalence criteria. Below we present instances of such pairs of LUs (for their full descriptions see Sections 3.1 and 3.2):

- ***narzeczona***$_{n:1}$ I-syn ***fiancee***$_{n:1}$
- ***centrala***$_{n:2}$ I-syn ***headquarters***$_{n:1}$
- ***gruba ryba***$_{n:1}$ I-partial-syn ***big fish***$_{n:1}$

  All strong because of identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

The last group of synsets to look at are many-to-many sense pairs, among which we are likely to find pairs of LUs that can function as strong equivalents of each other. These are illustrated below (for their full description see Section 1 and 3.2):

- ***złoto***$_{n:3}{}^{PL}$I-syn ***gold***$_{n:3}{}^{EN}$
- ***okulary***$_{n:1}{}^{PL}$I-syn ***glasses***$_{n:3}{}^{EN}$

  For all, identity in sense and register, frequent (often first) dictionary listing, many parallel corpora hits

4.2 Regular equivalence

The regular equivalence will be a slightly weaker type of link than the strong one, but it will still signal clear correspondence in a number of features. It will require large similarity in sense, compatibility in register, dictionary listing, bidirectionality of translation, a similar set of typical co-texts and, preferably, some parallel corpora hits. It will allow for some differences in lexicalisation of concepts. Examples of regular equivalence links from one-to-many sense pairs are given below (for their full descriptions see Section 3.2):

- ***zabytek***$_{n:1}$ I-partial-syn ***monument***$_{n:2}$

  Lexical gap (on the English side)

- ***narzeczona***$_{n:1}$ I-syn ***bride-to-be***$_{n:1}$

  Additional (temporal) sense specification on the English side; few parallel corpora hits

- ***centrala***$_{n:2}$ I-syn ***central office***$_{n:1}$

  Few parallel corpora hits for this pair

Instances of regular equivalence can also be found within many-to-many sense pairs. Below we illustrate them with instances of Polish grammaticalised gender (for their full description see Section 3.1) :

- ***nauczyciel***$_{n:1}$ I-syn ***teacher***$_{n:1}$
- ***nauczycielka***$_{n:1}$ I-hypo ***teacher***$_{n:1}$

  Examples of Polish grammaticalised gender

4.3 Weak equivalence

Since translatability can be achieved by very different means, we would like to point out that in certain contexts even LUs from pairs that do not meet all the criteria for strong or regular equivalence can function as translational equivalents. We will call such type of equivalence weak (or implied) equivalence. It will be postulated for pairs of LUs from plWordNet and Princeton WordNet synsets linked by I-synonymy, I-partial synonymy and I-hypernymy that do not meet the criteria for strong or regular equivalence, and can be automatically derived from the synset-level links. Often these will be instances of culture specific concepts absent from the second language (cultural gaps) and linked via I-hyponymy relation. An example of such weak equivalence link is given below. It obtains for both component LUs of the Polish synset given below (for its full description see Section 3.2.):

- {***stachanowiec***$_{n:1}$, ***przodownik pracy***$_{n:1}$} I-hypo {***toiler***$_{n:1}$}
  Polish culture specific term, with no direct equivalent

We expect that, except for instances of lexical gaps and gender lexicalisation where bidirectionality of translation does hold, the majority of I-hyponymy and I-hypernymy synset links will be unidirectional in terms of translation and thus pairs of their component LUs will be treated as weak equivalents.

5 Linking procedure

Having defined the equivalence features and types, below we put forward a linking procedure for lexicographers. In the procedure we lead lexicographers from simpler to more complex features and from wordnet data to dictionary and corpora data. We believe that there is no need for a lexicographer to verify each feature separately, but that they can be analysed in groups or pairs on the basis of the data provided by a specific resource.

We will illustrate the linking procedure on the example of the pair of synsets {***centrala***$_{n:2}$} linked via I-synonymy to {***headquarters***$_{n:1}$, ***central office***$_{n:1}$, ***main office***$_{n:1}$, ***home office***$_{n:2}$, ***home base***$_{n:2}$}. Formal features that is number, countability and gender should be verified first. Gender is not relevant here, since we do not deal with an animate noun. On the other hand, we have an instance of a pluralia tantum in the English synset: ***headquarters***$_{n:1}$. The remaining lexical units are regular countable nouns. Next, we move to semantic (and partly pragmatic) features starting from the data provided in wordnets that is relations, glosses, qualifiers and examples. The key relations are hypernyms and hyponyms, as well as their I-synonyms or I-hypernyms. The Polish synset {***centrala***$_{n:2}$} has {***ośrodek***$_{n:2}$, ...} - 'center' as its hypernym, which is an I-synonym of the English {***centre***$_{n:4}$, ...}. It is glossed as: " siedziba centrali, główny ośrodek czegoś" - ' the headquarters' seat, main centre of something'. It has general register and the usage example is the following: "Pożar centrali mleczarskiej w miejscowości obok było widać z daleka." - 'The fire in the dairy center in the nearby place could be seen from the distance.' The English synset {***headquarters***$_{n:1}$, ...} has {***office***$_{n:1}$, ***business office***$_{n:1}$} as its hypernym. It is attributed with the following gloss and example: "(usually plural) the office that serves as the administrative center of an enterprise; "many companies have their headquarters in New York." There is no information about the register provided.

Next, in order to gather still more information about semantics and pragmatics as well as translatability of pairs of particular LUs, lexicographers are asked to consult external resources such as dictionaries and encyclopedias as well as a Polish-English parallel corpus *Paralela*. Looking up *centrala* in a couple of Polish-English dictionaries (see ...), we find that its most frequent equivalents are *headquarters*, *head office* and *central office*. Interestingly, *head office* does not appear in Princeton WordNet at all. Looking up *headquarters* in English-Polish dictionaries, we obtain *centrala* and *siedziba główna* (the latter term appearing in the gloss of the Polish synset); checking *central office*, we get *siedziba główna* and *centrala*. In the next step, we check the frequency of the pairs *centrala – headquarters* and *centrala – central office* in the *Paralela* corpus and we learn that the pair *centrala* and *headquarters* gets 40 hits, while *centrala – central office* gets only 3 hits. In the last step, we analyse the most frequent contexts of occurrence of *centrala – headquarters* and we get a couple of typical shared cotexts and collocations (examples given in Section 3.2.) On the basis of the whole discussed data, we want to argue that the lexical units ***centrala***$_{n:2}$ -

headquarters$_{n:1}$ form a pair of strong equivalents, ***centrala***$_{n:2}$ - ***central office***$_{n:1}$ are regular equivalents, while ***centrala***$_{n:2}$ - ***main office***$_{n:1}$, ***home office***$_{n:2}$, ***home base***$_{n:2}$ should be treated as weak equivalents.

6 Conclusions

The strategy for sense-level mapping between Princeton WordNet and plWordNet nouns put forward in this paper is a new initiative in the wordnet world. It offers a possibility for fine-grained mapping that is of great potential especially for human and machine translation. It is illustrated with examples from the Polish-English language pair, but the set of features described in this paper are language-neutral and they can be easily extended to wordnets of other languages of the Indo-European family. As for (non)-Indo-European language pairs, it is necessary to analyse whether the two languages share all the features that will be taken into account. Also, the strategy may be extended to other grammatical categories such as adjectives and adverbs, which are already partially mapped on the synset level, and, eventually, to verbs after some mapping between verb synsets is accomplished. It may well be that additional features will need to be introduced while some of the ones proposed for nouns might be dismissed as irrelevant.

The proposed strategy is designed for manual mapping, but we plan to develop an automatic system of prompts that will support lexicographers' work. The new system will be an extension of an earlier system of automatic prompts for mapping of noun synsets and based on a modification of the Relaxation Labelling algorithm of Daudé et al. (1999) joined with lemma-pair checking and filtering by a large Polish-English cascade dictionary Kędzia et al. (2013) and translation probabilities from bilingual corpora.

As regards future avenues, this study may be continued in a number of possible ways. Firstly, the strategy of sense-level mapping described in this paper should be further tested on a structured and balanced sample of concrete and abstract nouns representing the whole variety of semantic domains (lexicographers' files). We plan to extract the lists of Polish-English lexical unit pairs from the Polish-English pairs of synsets linked by I-synonymy, I-partial synonymy and I-hyponymy (both Polish-English and English-Polish). The reason for that is that pairs linked by these relations are most likely to yield strong and regular equivalents. We will (proportionally) explore all three possible types of pairing, that is 1-to-1 sense match, 1-to-many sense match and many-to-many sense match.

Secondly, in order to pinpoint any translation tendencies, the next step should be to calculate translation probabilities for pairs of equivalents, preferably in both directions, extracted from parallel corpora (e.g. *Paralela*). This would enable the verification of the degree to which sense-level mapping is reflected in translated texts found in a parallel corpus. Obviously enough, translation probabilities should be interpreted with caution given the limitations of any parallel corpus used (its size, structure, representativeness, balance, scope of annotation, etc.). At this point, it is also important to note that searching through parallel corpora is problematic when one deals with polysemous lexical units. The lack of word-sense disambiguation (or, in other words, semantic tagging of bilingual corpus data) means that when we consult a parallel corpus, we search for language forms rather than senses; that is why translation probabilities should be calculated in a way reflecting polysemy of lexical units. All this should enable one to further test, verify and improve the linking procedure proposed in this paper, which can be useful for anyone interested in applying it for sense-level mapping of wordnets representing languages other than Polish and English.

Acknowledgment

This work was supported by the National Science Centre in Poland under the agreement No. UMO-2015-/18/M/HS2/00100.

References

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

Luisa Bentivogli and Emanuele Pianta. 2004. Extending wordnet with syntagmatic information. In *Proceedings of the Second Global WordNet Conference*, pages 47–53. Brno, Czech Republic.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proc. 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

CALD. 2013. *Cambridge Advanced Learner's Dictionary*. Cambridge University Press, Cambridge, fourth edition.

Cobuild. 2012. *Collins Cobuild Advanced Dictionary of English.* Heinle ELT, Boston, 7th edition.

Collins-YDP. 1997. *Multimedialny słownik angielsko-polski i polsko-angielski Collins*. Polska Oficyna Wydawnicza, Warszawa. Version 3.03 (computer program).

Jordi Daudé, Lluís Padró, and German Rigau. 1999. Mapping multilingual hierarchies using relaxation labeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 12–19.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. The MIT Press.

Eniko Heja. 2016. Revisiting translational equivalence: Contributions from data-driven bilingual lexicography. *International Journal of Lexicography*.

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

LDCE. 2014. *Longman Dictionary of Contemporary English*. Pearson Education, Harlow, 6th edition.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNorcdica*, 17.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plWordNet as the cornerstone of a toolkit of lexico-semantic resources. In Heili Orav, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312. University of Tartu Press, Tartu, Estonia. URL `http://aclweb.org/anthology/W14-0142`.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL. URL `http://aclweb.org/anthology/C/C16/`.

PWN-Oxford. 2007. *Wielki słownik angielsko-polski PWN-Oxford*. Wydawnictwo Naukowe PWN S.A. and Oxford University Press, Warszawa.

Anthony Pym. 2007. Natural and directional equivalence in theories of translation". *Target*, 19(2):271–294.

Piotr Pęzik. 2016. Exploring phraseological equivalence with paralela. In Ewa Gruszczyńska and Agnieszka Leńko-Szymańska, editors, *Polish-Language Parallel Corpora*, pages 67–81. Instytut Lingwistyki Stosowanej UW, Warszawa.

Ewa Rudnicka, Francis Bond, Łukasz Grabowski, Maciej Piasecki, and Tadeusz Piotrowski. 2017a. Towards equivalence links between senses in plWordNet and Princeton WordNet. *Lodz Papers in Pragmatics*, 13(1):3–24.

Ewa Rudnicka, Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. A Strategy of Mapping Polish WordNet onto Princeton WordNet. In *Proc. COLING 2012, posters*, pages 1039–1048.

Ewa Rudnicka, Maciej Piasecki, Tadeusz Piotrowski, Łukasz Grabowski, and Francis Bond. 2017b. Mapping wordnets from the perspective of inter-lingual equivalence. *Cognitive Studies / Études cognitives*, 17. In print.

Ewa Rudnicka, Wojciech Witkowski, and 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 B. Barbu Mititelu, C. Forascu, Ch. Fellbaum, and P. Vossen, editors, *Proceedings of the 8th International Global WordNet Conference 2016,*, pages 344–351. Global WordNet Association, Bucharest, Romania. URL `http://gwc2016.racai.ro/procedings.pdf`.

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

Słownik-Kościuszkowski. 2014. *Nowy słownik Fundacji Kościuszkowskiej polsko-angielski i angielsko-polski*. TAiWPN Universitas, Kraków.

Techland-Dictionary. 2006. *Wielki słownik angielsko-polski, polsko-angielski*. Techland. Version 1.0.1 (computer program).

Piek Vossen. 2002. EuroWordNet General Document Version 3. Technical report, Univ. of Amsterdam.

Ladislav Zgusta. 1971. *Manual of Lexicography*, volume 39 of *Series Maior*. Janua Linguarum, The Hague. Pari.