Non-Lexicalized Concepts in Wordnets: A Case Study of English and Hungarian

Veronika Vincze[1,2]
[1]MTA-SZTE Research Group on Artificial Intelligence
[2]University of Szeged
vinczev@inf.u-szeged.hu

Attila Almási
University of Szeged
almasia@inf.u-szeged.hu

Abstract

Here, we investigate non-lexicalized synsets found in the Hungarian wordnet, and compare them to the English one, in the context of wordnet building principles. We propose some strategies that may be used to overcome difficulties concerning non-lexicalized synsets in wordnets constructed using the expand method. It is shown that the merge model could also have been applied to Hungarian, and with the help of the above-mentioned strategies, a wordnet based on the expand model can be transformed into a wordnet similar to that constructed with the merge model.

1 Introduction

Wordnets are lexical databases in which words are organized into clusters based on their meanings, and they are linked to each other through different semantic and lexical relations, yielding a conceptual hierarchy (i.e. lexical ontology) of words. Originally, they were designed to show how linguistic knowledge is organized within the human mind (Miller et al., 1990). Multilinguality is also an important aspect in the creation of wordnets: builders of new wordnets usually map their synsets to those representing the same concept in Princeton WordNet (PWN).

However, there is no perfect mapping between two languages at the conceptual level and the lexical level. In this article, we would like to compare the wordnets built for Hungarian and English and we will discuss problems and possible solutions concerning discrepancies in the way the two languages name certain concepts in the context of wordnet-building methods and principles. First, the wordnets we study are briefly presented, then the notions of non-lexicalized and technical non-lexicalized synsets are illustrated with concrete examples. We suggest some ways of eliminating non-lexicalized synsets from wordnets, and we also show how a Hungarian tree can be built without relying on the English tree. Lastly, we argue that although a wordnet that seeks to represent the hierarchy of the given language should not contain non-lexicalized elements, they can prove useful in fields of research such as psycholinguistics, ethnography and contrastive linguistics.

2 Related Work

The first wordnet was created for the English language at Princeton University, so it is called the Princeton WordNet (Fellbaum, 1998). It is now the largest lexical database of the English language, and it can be readily adapted to various computational applications. Princeton WordNet 3.0 contains about 155,000 words in approximately 117,000 synsets.

Since then, other wordnets have been created and developed for different languages. EuroWordNet is a multilingual project, where synsets for Dutch, Italian, Spanish, German, French, Czech and Estonian are included in the database (Alonge et al., 1998). The BalkaNet project sought to extend EuroWordNet with lexical databases created for languages of the Balkan Peninsula, namely Bulgarian, Greek, Turkish, Serbian and Romanian (Tufiş, 2004; Tufiş et al., 2004). Other languages for which wordnets have been developed include Arabic, Croatian, Chinese, Danish, Slovene, Polish, Russian, Persian, Hindi, Tulu, Dravidian, Tamil, Telugu, Sanskrit, Assamese, Filipino, Gujarati, Nepali (Tanács et al., 2008; Bhattacharyya et al., 2010; Fellbaum and Vossen, 2012).

Typically, there are two major approaches to wordnet construction (Vossen, 1998). The first approach (merge model) starts by constructing a wordnet from scratch (or by using dictionaries and other resources developed for the language) and then the newly created synsets are linked to synsets of another language (most typically English). The second approach (expand model) starts by selecting a subset of the PWN synsets and then they are transformed into synsets of the target language, preserving relations between

synsets. Wordnets created in this way inevitably reflect lexicalization of the given language to a lesser degree; however, it is known that the nodes in PWN form a network, the rendering of which into the given language may be unnatural, forced and this may result in further difficulties concerning multilingual applications (Raffaelli et al., 2008). The merge model was used for most languages in the EuroWordNet project (Alonge et al., 1998), whereas the expand model was used for Spanish, Hungarian and some other languages.

Now, languages do not overlap completely: due to the differences in culture, traditions and lifestyle, languages have concepts, words characteristic of the given language alone. They can only have approximate equivalents and cannot be translated using a single word (Derwojedowa et al., 2008), i.e. they cannot be lexicalized.

Lexicalization is defined in the following way (Lipka, 1992: 107): "the process by which complex lexemes tend to become a single unit with a specific content, through frequent use. In this process, they lose their nature as a syntagma, or combination, to a greater or lesser extent." Thus, lexicalization can be regarded as a process that is gradual, similar to the scalar view of productivity (Jackendoff, 2010). Thus, there are lexicalized items in the language, there are non-lexicalized ones and there are borderline cases in between.

For non-lexicalized concepts, artificial nodes may be introduced in wordnets so as to have a better organized structure (Fellbaum, 1998). The original PWN also contains a few such items, e.g. *bad person*. However, there are wordnets which contain only lexicalized concepts of a language and no non-lexicalized synsets are included. For instance, the Dutch wordnet does not include artificial synsets, producing a much flatter hierarchy (Vossen, 1998). Despite this, the creators of the Basque wordnet tried to include as many non-lexicalized multiword expressions as possible (Agirre et al., 2006). They differentiate between conceptual level imbalances and expression level imbalances, similar to Vossen (1999), who distinguishes cultural gaps and pragmatic gaps. The Basque wordnet, which was also built following the expand model, explicitly codes these non-lexicalized synsets (Pociello et al., 2011).

The Hungarian WordNet (HuWN) was developed by the Research Institute for Linguistics of the Hungarian Academy of Sciences, the Department of Informatics of the University of Szeged, and MorphoLogic Ltd. in a 3-year project (Alexin et al., 2006; Miháltz et al., 2008). As a result, HuWN now contains over 40,000 synsets, out of which 2,000 synsets form part of a business subontology. Here, Princeton WordNet 2.0 served as a basis for the construction of HuWN, i.e. the expand model was adhered to. More precisely, synsets belonging to the BalkaNet Concept Set were selected from PWN 2.0 and then translated into Hungarian. These were then manually edited, corrected and extended with other synonyms using the VisDic editor. The set of concepts to be included in HuWN were expanded concentrically later on. That is, descendants of the existing synsets were treated as synset candidates. The final decision on their status (whether they should be included or excluded) was influenced by several factors such as the frequency of the concept and its presence in other WordNets (Miháltz et al., 2008).

In this paper, we examine what the effects of the expand model are on the quality of the Hungarian WordNet. We investigate the types of non-lexicalized synsets and we propose some strategies that may be used to overcome difficulties concerning non-lexicalized synsets in wordnets constructed using the expand method.

3 Non-Lexicalized Synsets

At its inception, developers of the Hungarian wordnet decided that the so-called expand method should be used. This implies that HuWN inherited the hierarchy of PWN. The nominal and adjectival parts[1] of HuWN were built according to the following method: nodes in PWN were automatically correlated with Hungarian synsets and their relations were adopted; the basic strategy was to attach Hungarian entries of a bilingual English-Hungarian dictionary to the nominal/adjectival synsets of PrincetonWordNet.

In order not to have "holes" in the constructed tree (that is, in order for the English and Hungarian wordnets to overlap as much as possible), developers had to find a good way of handling such synsets. To indicate that such synsets do not exist (at the word level) in the lexicon of the given language, i.e. they have not become lexicalized, the *non-lex* label was introduced. Now, we will give the criteria for a synset to be non-lexicalized. First, it may be that no such concept exists in the given language (especially due to cultural differences). Second, the concept may be

[1] The verbal part of HuWN was constructed in a different way (cf. Kuti et al., 2008), so we did not consider verbs in our study.

expressed by productive and compositional constructions (e.g. with adjective + noun combinations), i.e. there is no way of expressing it using a single word or a multiword expression. Third, the concept may be an umbrella term for several single-word concepts, thus, in the other language it may only be expressed by a list. Fourth, there seemed to be inconsistencies or erroneous definitions and hypernym relations in PWN, which the builders of the Hungarian wordnet did not want to follow and they marked the problematic synset with the *non-lex* label.

Some statistics on non-lex synsets in HuWN are presented in Table 1. It can be seen that for the whole body of HuWN every twentieth synset is non-lexicalized and for the basic concept set (BCSHu) it is every twelfth synset. Hence, the problem is not negligible and it is worth examining in detail what types of nonlex synsets exist and how they can be eliminated.

	HuWN	**BCSHu**
Synsets	42,292	8446
Non-lexicalized	1,999	463
Technical non-lexicalized	454	271
% of (t)non-lex synsets	5.799	8.69

Table 1: (Technical) non-lex synsets in HuWN.

3.1 Types of Non-Lex Synsets

Non-lex synsets found in HuWN can be classified into six main groups, which are presented below.

Culturally Determined Concepts. Culturally determined concepts are related to differences in culture, lifestyle or geographical background. Since the American and Hungarian cultures, (folk) traditions and backgrounds are quite different, there are concepts which not always have verbatim equivalents in the other language. In case they have, they may not reflect the feelings and moods they evoke, that is, what comes to a person‘s mind when he hears them may differ in the two cultures (cf. Zidoum, 2008). Here we provide two examples:

máglyarakás ‘stake’ (in Hungarian, it refers to a kind of confectionery, which is not associated with the English word *stake*).

Sassenach – a Scot’s term for an English person, where connotations of the original word cannot be mirrored in Hungarian.

Culturally determined concepts are called conceptual level imbalances in the Basque wordnet (Pociello et al., 2011).

Geographical background mostly determines the named entities included in wordnets. For instance, most Hungarian speakers are not familiar with **Milk River:1** or **White River:1**, thus their inclusion would be questionable in the Hungarian wordnet. However, some of them are included in HuWN due to the expand method applied, but they are classed as *non-lex*.

Split Concepts. Another group of non-lex synsets includes elements that simply have no counterpart in the given language. Very often, certain umbrella terms belonging to this category can only be expressed in the other language by using a paraphrase or supplying a list. For instance, **cycling:1** is used for both riding bicycles and motorcycles, which are separate lexical units in Hungarian.

Words with a Negative Prefix. Another basic example of non-lex synsets is that of adjectives/nouns formed with negative prefixes such as *non-*, *in-* and *un-*. Apart from a couple of cases, in Hungarian, the negated version of such lexical units is produced with a negative adverb and they together do not constitute a lexicalized synset. Examples of non-lex synsets in HuWN formed with negative prefixes in PWN include **unattractive** – *nem vonzó*, **ill-timed** – *rosszul időzített* and **incongruity** – *meg nem egyezés*, where the HuWn synsets are marked as non-lexicalized.

Adjective + Noun Constructions. Some concepts in PWN are expressed with adjective + noun constructions in Hungarian, which cannot be regarded as lexicalized units since they are productive and their meaning is totally compositional. For instance, words denoting nationalities (*skót* ‘Scottish’, *angol* ‘English’, *magyar* ‘Hungarian’ etc.) in Hungarian have a peculiar feature that although there is no distinction of gender in the nominal and pronominal system at the morphological and syntactic levels, when using these words we first and foremost mean a male person of a nation: e.g. **Scotsman:1** was annotated *skót* (a Scottish male person). Their female counterpart is usually formed by adding an extra noun, *nő* ‘woman’. The two words *skót nő* ‘Scottish woman’ when combined, however, are regarded as a productive construction (of adjective + noun) and not as a multiword expression, which is a prerequisite for Hungarian adjective + noun constructions to be admitted into HuWN as valid synsets, and hence *skót nő* is a non-lexicalized synset paired with **Scotswoman:1, Scotchwoman:1**.

Linguistic Differences. Sometimes non-lexicalized synsets arise due to the ways a concept can be expressed. In the case of **people:1** – (embercsoport), it can be expressed by a suffix in Hungarian: the English phrase *200 people* can translated as *kétszázan* two.hundred-ESSIVE into Hungarian, which means that a suffix denoting the essive grammatical case is attached to the number, and the suffix corresponds to the English noun.

Technical Terms. Over the course of time, some non-lexicalized concepts may become lexicalized. One typical domain is technology, where such concepts are spreading worldwide at an ever accelerating rate. A few years ago, when HuWN was being constructed, *RV* (recreational vehicle) for instance was tagged *non-lex*, which, now, could be accepted as a fully acknowledged lexicalized synset.

3.2 Technical Non-Lexicalized Synsets

During the construction, it frequently happened that two English synsets in hierarchical relation had a single Hungarian equivalent; the two concepts are distinct at the conceptual level only. At the lexical level, however, it is impossible to find two distinct words for them. In other cases, it was not possible to find an equivalent for the word with the same part of speech. Technical non-lexicalized (t non-lex) tags are applied in the following cases: (1) identical literals in hypernym-hyponym relation; (2) identical literal in a *similar_to* relation; (3) POS difference, which are all illustrated below.

Identical Literals in Hypernymy Relation. The first case of technically non-lexicalized tagging in HuWN is when there are two identical literals in synsets in hypernym relation. This phenomenon is called autohyponimy in Cruse (2000). The developers of HuWN wanted to avoid such redundancies in the trees and, as a convention, they eliminated the overlapping literal from one of the synsets.

Due to entailment, a concept can be replaced by its hypernym: if a greyhound barks, then it entails that a dog barks. So it seemed reasonable to apply this axiom in HuWN building, i.e. to not repeat the hypernym in the hyponym synset. Here is an example (the numbers denoting levels of hierarchy):

1 **cube:5**	**kocka:3**
2 **dice:1**	**dobókocka:1**

In this case, due to the above-mentioned convention of having to delete the identical literal in the hyponym synset, *kocka* has been excluded, leaving only *dobókocka* as a hyponym. Thus, there is no need to mark the hyponym synset as technically non-lexicalized since there is another literal which does not coincide with the hypernym.

In cases where the hyponym synset consists of only one literal, coinciding with its hypernym, the hyponym synset is marked t non-lex:

1 **safety:1**	**biztonság:1**
2 **security:1**	**biztonság:0**

In Hungarian, there is no separate lexical item for *safety* and *security*, these being roughly equivalent to *biztonság*. In this way, the hyponym synset should be marked as t non-lex.

Identical Literals in Focal-Satellite Synsets. In the case of the adjectival part of the ontology, the *t non-lex* label was also employed. Since its construction is based on antonym-pairs and the associated, synonymous "satellite" synsets, it may well be that while distinct words in English are used to express the concept belonging to the focal and the satellite synsets, in Hungarian, the same word occurs in both positions. Yet, the conventions of wordnet building require that the focal and the satellite synsets should contain no identical literals (cf. identity of hypernym and hyponym). Consequently, again, the course to be followed is that the focal synset remains lexicalized and the more specific, satellite synset gets the *t non-lex* label. For example, {**wide:1**; **broad:1**}'s "satellite" synset is {**heavy:5**; **thick:5**}, but in Hungarian *széles* corresponds to both, therefore the focal synset will be {**széles:2**}, and the satellite synset {**széles:0**}.

Different Parts of Speech. Sometimes the target language equivalent of a synset does not share its part of speech with the source language word although it can be classified as one of the four parts of speech used in wordnets. For instance, the English word *afraid* is an adjective, but its Hungarian counterpart *fél* is a verb. In such cases, we made use of the relation *eq_xpos_synonym*, which designates synonymy among different parts of speech: here it relates *fél* and the Hungarian adjectival synset corresponding to *afraid*, which is marked as *t non-lex*.

4 Wordnet Errors Related to Non-Lexicalized Synsets

Now we present some of the problematic synsets from PWN and HuWN along with their solutions.

4.1 Problems in the Tree

In certain cases, there is an incongruence between a synset and its hypernym. For instance, **location:1** in PWN is defined as *a point or extent in space*; one of its hyponyms is **bilocation:1** with the definition of *the ability (said of certain Roman Catholic saints) to exist simultaneously in two locations* (unique beginner synset: **entity:1**). To our mind, this relation is invalid as their definitions are incompatible and only seem to make a formal hyper-hyponym pair. Instead, *bilocation* should be linked to **ability:2, power:3/képesség:2** on the basis of the definition given in PWN, or it could be also linked to **phenomenon:1/jelenség:1**. If the structure of PWN is to be preserved in HuWN, this synset should be marked as *non-lex* and a new synset should be created under the correct hypernym (**képesség:2** or **jelenség:1**).

4.2 Lexicalized Synsets Marked as Non-Lex

In our opinion, in certain cases the annotators of HuWN made some mistakes. For instance, **labor:1** is now a non-lex synset but it should have been classed as a full-fledged lexicalized synset, a multiword expression *fizikai munka* 'physical work'. Similarly, we think that **seating:1**, **area:1** should have been included as *ülőhely* 'seat'.

4.3 Non-Lexicalized Synsets Marked as Lexicalized

An interesting example of non-lex synsets is **bow and arrow:1/íj és nyílvessző:1**. In our view, the synset was incorrectly tagged lexicalized as – though the two parts make up a single weapon – the projector (bow) and the projectile (arrow) do not form a lexicalized phrase in Hungarian.

Attempts to find a Hungarian equivalent for PWN synsets sometimes led to such completely non-existent (although possible) synsets in Hungarian as **fúvóeszköz:1 (blower:1**).

5 Eliminating Non-Lex Problems

The large number of non-lexicalized synsets in the Hungarian wordnet raises questions concerning the (organizing) principles of the Hungarian wordnet. Non-lex synsets – strictly speaking – are not part of the given language, and wordnets including many non-lexicalized items can hardly be regarded as reflecting the concepts of the given language. In order to overcome these problems, we propose to minimize the number of non-lexicalized synsets with the help of four strategies, which are presented below.

5.1 Shortening the Tree

We suggest that non-lex synsets without any hyponym should be deleted from the tree. As hypernyms can substitute hyponyms in every context (see Section 3.2.1), this strategy does not undermine the expressibility of certain concepts. This might be useful in the following trees:

1	**freedom:1**	**szabadság:1**
2	**liberty:1**	(szabadság)

There is no distinction made between the senses of the PWN concepts in Hungarian, thus, the lower non-lex synset should be deleted. This solution may be applied to certain culture- or geography-specific synsets as well. For instance, it proved sufficient to include only the major rivers of the United States in HuWN, as there was no need to adapt all the rivers listed in PWN.

5.2 Flattening the Tree

Split concepts that can be paraphrased by giving a list should simply be deleted from the tree and all of their hyponyms can be attached to the hypernym of the deleted synset. For instance, there are two non-lex synsets in the following tree:

1	**occupation:1, business:6, job:1, line of work:1, line:19**	**foglalkozás:1, munka:3, hivatás:2, pálya:6**
2	**profession:2**	(foglalkozás)
3	**learned profession:1**	(jog, orvostan és hittudomány)
4	**law:5, practice of law:1**	**jog:2, jogtudomány:1**
	medicine:3, practice of medicine:1	**orvostudomány:1**
	theology:3	**hittudomány:1**

The first non-lex synset corresponds to the same lexical item as its hypernym in Hungarian, so it is unnecessary to include the non-lex synset in the Hungarian wordnet. The second non-lex synset corresponds to an umbrella term in English, which has no proper Hungarian counterpart. Instead, the following tree should reflect the real conceptual hierarchy in Hungarian:

1	**foglalkozás:1,munka:3, hivatás:2, pálya:6**
2	**jog:2, jogtudomány:1** **orvostudomány:1** **hittudomány:1**

5.3 Restructuring the Tree

In certain cases, the reconstruction of the tree may be the most effective. First of all, let us illustrate the problem with two charts representing the corresponding PWN and HuWN tree-sections (Hungarian paraphrases are equivalent to PWN definitions):

1	**building:1**	**épület:1**
2	**place of worship:1**	(istentisztelet helye "place of worship")
3	**church:2**	(keresztény templom "Christian church")
	temple:1	(nem keresztény templom "non-Christian church")

In PWN, **church:2** and **temple:1** are hyponym synsets of **place of worship:1** at the same level while, at present, they have no lexicalized counterparts in the Hungarian wordnet. In order to eliminate the three non-lexicalized synsets in HuWN and to have lexicalized items there, we propose a solution in which *templom* (meaning a building for the worship of any deity or any religion in Hungarian, without distinguishing between a Christian or non-Christian place of worship) is placed in the hypernym position in parallel with **place of worship:1** and the two hyponym synsets in PWN have no counterparts in the Hungarian tree. All the original hyponyms of **church** and **temple** can be linked under **templom** in Hungarian now.

1	**building:1**	**épület:1**
2	**place of worship:1**	**templom:1**
3	**church:2**	(-)
	temple:1	(-)

5.4 Lexicalizing the Concept

In some cases, it happened that wordnet builders had made an error and marked lexicalized concepts as non-lex (see Section 4.2). In other cases (see Section 3.1.6), certain concepts (mostly from the technological domain) became lexicalized over time and now they are genuine members of the Hungarian language. The non-lex label of these synsets should be deleted and the synset should be treated as lexicalized, i.e. providing the definition, usage and literals for it.

6 Building Independent Hungarian Trees

At the outset of the project, wordnet builders decided to follow the expand model, which meant that HuWN was largely built by simply translating PWN synsets and taking over its relations. To test the validity of this decision, we experimented with the merge model and we also built trees that are truly representative of the structure of the Hungarian language so as to compare Hungarian and English trees.

Hence, we decided to build an independent Hungarian tree from scratch and to examine if we could find matches in HuWN and PWN. First, we took a brand of the famous Hungarian wine called Tokay aszu. The following chart illustrates the newly constructed Hungarian and the corresponding English tree from the top down. [m*X*] denotes synsets that make perfect matches in the independent Hungarian tree, HuWn and PWN. At level 8, there are two relevant concepts that are hyponyms of *fehérbor*. *Tokaji aszú* at level 10 is a hyponym of both *aszúbor* and *tokaji*.

1	entitás:1	[m7] *entity*
2	anyag:1	[m6] *substance*
3	folyadék:2 | tápanyag:1	[m5] *liquid* | *food*
4	ital:1	[m4] *beverage*
5	szeszes ital:1	[m3] *alcohol*
6	bor:1	[m2] *wine*
7	fehérbor:1	[m1] *white wine*
8	desszertbor | tokaji	*dessert wine* | *Tokaji*
9	aszúbor	*aszu wine (botrytized wine)*
10	tokaji aszú (hyponym of *tokaji* too)	*aszu wine from Tokaj*
11	hatputtonyos tokaji aszú	*six-puttonyos Tokay aszu*
12	Oremus hatputtonyos tokaji aszú	*six-puttonyos Tokay aszu from Oremus winery*

Concepts at levels 9-12 cannot be found in HuWN at all and have no corresponding synsets in PWN either. The concepts at level 8 have no corresponding synsets in HuWN, however, *desszertbor* has a lexical and conceptual counterpart in PWN.

There seems to be a problem regarding the concept *tokaji* in the above chart and the synset *Tokay* in PWN. *Tokaji* in Hungarian (and in Eng-

lish language sources as well[2]) refers to all the wines produced in the Tokaj district of North-eastern Hungary. This concept does not seem to have an equivalent in PWN: it certainly has no formal equivalent and it cannot be decided what the definition of the synset **Tokay:1** (PWN definition: Hungarian wine made from Tokay grapes) refers to exactly. To our mind, it seems closer in meaning to Tokay aszu, which was formerly known throughout the English-speaking world as Tokay (Webster's 1913). Thus, it seems that the Hungarian concept, *tokaji* – which was not included in HuWN – has no equivalent in PWN.

Fehérbor (*white wine*) splits into *desszertbor* (*dessert wine*) and *tokaji* (*Tokaji*) at level 8, only to merge again at *tokaji aszú* (*Tokay* (*aszu*)), at level 10. *Aszúbor* (*botrytized wine*) at level 9 is a non-existent synset in PWN.

The tree was built from scratch but it is quite evident that – apart from the levels below 7 – it matches perfectly the Hungarian wordnet: synset numbers are actual sense numbers found in HuWN. **Ital:1** has two hypernyms, both merging into the same hypernym at level 2. These facts suggest that a merge model would also have been applied in the construction of HuWN.

7 Discussion

Since languages and cultures differ from each other, there are necessarily concepts that may be lexicalized in one but not in the other and vice versa. Non-lexicalized elements reflect either conceptual or cultural differences between languages and hence can be used for checking the similarities among languages. The Hungarian wordnet – having been constructed according to the expand model – in its present form contains a relatively high number of non-lexicalized synsets but should there be a revision, they might be deleted from the tree (either by shortening or flattening the tree), the tree might be restructured, or they might be lexicalized (if erroneously annotated as *non-lex*). In this way, the Hungarian wordnet would really reflect the hierarchy of the Hungarian language.

Our experiments with building independent Hungarian trees showed that it would also have been viable to apply the merge model for wordnet building. Most of the synsets within the trees can be linked to a corresponding English synset, thus, interlinguality can also be assured as well.

The results of our experiments also led us to ask whether it was justifiable to include non-lexicalized items in PWN. From a purely lexical point of view, these concepts do not exist in the language and so may be deleted from the hierarchy. The argument that should there be no *good person* and *bad person* synsets in PWN, *offender* and *lover* would be sisters, being the hyponyms of *person* (Fellbaum 1998) can be refuted by stating that this would not cause much difficulty given that among the children of *person*, we can already find synsets denoting positive concepts (*enjoyer*), negative concepts (*killer*) and neutral concepts (*candidate*). A second issue concerning PWN is that although it was intended to model the human mind, there are concepts that cannot be found there: see the example of elder and younger brothers and sisters, which are separate lexical items in Hungarian, so they denote different concepts and if the original plan had been followed, they should occur in PWN too – at least as non-lexicalized synsets. A third issue with PWN is that no distinction is made between lexicalized and non-lexicalized ones, i.e. no labels like *non-lex* are used, which somewhat undermines its usage as a dictionary. Although PWN was intended to reflect the hierarchy of concepts thought to be universal, it is very often used as a traditional dictionary of lexical units and hence it should be the case that lexicalized and non-lexicalized concepts are distinguished.

In spite of this, we argue that the marking of non-lex synsets can be profitable as well, especially in an interlingual context. Researchers from different fields can exploit the benefits of non-lex synsets. Psycholinguists might want to compare the hierarchy of mental concepts of speakers of different languages – with the help of non-lex labels since differences are explicitly marked in wordnets built using the expand method. Culture-specific non-lex synsets might be used in ethnographic research. Non-lex synsets associated with linguistic differences (e.g. negative prefixes) can contribute to theoretical linguistic research and contrastive linguistics.

Based on the above points, we may conclude that the usability of wordnets is greatly influenced by the way they were constructed. Wordnets based on the merge model match the lexical hierarchy of the given language, so they can be used as dictionaries as well and they do not include marked non-lexicalized synsets. Due to the absence of non-lex synsets, matching them to other languages is quite difficult and they can be used for psycholinguistic comparative studies

[2] http://en.wikipedia.org/wiki/Tokaji

only in a limited way. Wordnets based on the expand model – such as HuWN – mainly follow the conceptual hierarchy defined in PWN, and contain a lot of non-lexicalized synsets. They can be used for making interlingual or psycholinguistic comparisons, but they reflect the structure of the given language to a lesser degree. However, with the strategies of deleting unnecessary non-lex synsets and restructuring the tree, it is possible to eliminate some of the non-lexicalized items and the wordnet based on the expand model may gradually converge to the one based on the merge model, without involving the effort of building a new wordnet from scratch.

8 Summary

In this study, we examined the precise effects of the expand model on the quality of the Hungarian WordNet. We investigated the types of non-lexicalized synsets and we proposed some strategies – including deleting superfluous synsets and reorganizing the trees – that may be used to overcome difficulties concerning non-lexicalized synsets in wordnets constructed with the expand method. We also presented an independent Hungarian tree – built to reflect Hungarian hierarchy and concepts – to see whether we could find matches with HuWN and PWN. It was shown that the merge model could also have been applied to Hungarian, and with the help of the above-mentioned strategies, a wordnet based on the expand model can be transformed to a wordnet similar to the one constructed with the merge model, which would reflect the conceptual hierarchy of Hungarian better. As the way of construction strongly influences the usability of wordnets, this latter version can be primarily used in intralingual research that focuses on Hungarian. Still, marked non-lexicalized elements could prove useful in different fields of research such as psycholinguistics, ethnography and contrastive linguistics. Hence, the originally published version based on the expand model can be also utilized in different fields of research.

In the future, we would like to modify the Hungarian wordnet and by eliminating superfluous non-lexicalized items, we would like to develop a wordnet that really takes into account the Hungarian way of lexicalizing mental concepts.

Acknowledgments

This work was in part supported by the European Union and co-funded by the European Social Fund through the project Telemedicine-focused research activities in the fields of mathematics, informatics and medical sciences (grant no.: TÁMOP-4.2.2.A-11/1/KONV-2012-0073).

References

Zoltán Alexin, János Csirik, András Kocsor, Márton Miháltz, and György Szarvas. 2006. Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In *Proceedings of GWC 2006*, pages 291–292, South Jeju Island, Korea.

Eneko Agirre, Izaskun Aldezabal, and Elisabete Pociello. 2006. Lexicalization and multiword expressions in the Basque WordNet. In *Proceedings of the Third International WordNet Conference (GWC2006)*, pages 131–138, Jeju Island, Korea.

Antonietta Alonge, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellon, Maria Antonia Marti, and Wim Peters. 1998. The Linguistic Design of the EuroWordNet Database. *Computers and the Humanities. Special Issue on EuroWordNet*, 32(2–3): 91–115.

Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors. 2010. *Principles, Construction and Application of Multilingual Wordnets. Proceedings of GWC 2010*. Mumbai, India, Narosa Publishing House.

Alan Cruse. 2000. *Meaning in language: An introduction to semantics and pragmatics*. London, Oxford University Press.

Magdalena Derwojedowa, Maciej Piasecki, Stanisław Szpakowicz, Magdalena Zawisławska, and Bartosz Broda. 2008. Words, Concepts and Relations in the Construction of Polish WordNet. In *Proceedings of GWC 2008*, pages 167–68, Szeged, University of Szeged, Department of Informatics.

Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Christiane Fellbaum, and Piek Vossen, editors. 2012. *Proceedings of GWC 2012*. Matsue, Japan.

Ray Jackendoff. 2010. *Meaning and the Lexicon: The Parallel Architecture 1975–2010*. Oxford University Press, Oxford.

Judit Kuti, Károly Varasdi, Ágnes Gyarmati, and Péter Vajda. 2008. Language Independent and Language Dependent Innovations in the Hungarian WordNet. In *Proceedings of GWC 2008*, pages 254–268, Szeged, University of Szeged, Department of Informatics.

Leonhard Lipka. 1992. Lexicalization and institutionalization in English and German. Or: Piefke, Wendehals, smog, perestroika, AIDS etc. *Zeitschrift für Anglistik und Amerikanistik* 40:101–111.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In *Proceedings of GWC 2008*, pages 311–320, Szeged, University of Szeged, Department of Informatics.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: an On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45:121–142.

Ida Raffaelli, Marko Tadić, Božo Bekavac, and Željko Agić. 2008. Building Croatian WordNet. In *Proceedings of GWC 2008*, pages 349–359, Szeged, University of Szeged, Department of Informatics.

Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors. 2008. *Proceedings of GWC 2008*. Szeged, University of Szeged, Department of Informatics.

Dan Tufiş, editor. 2004. *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, 7(1–2).

Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, 7(1–2):9–43.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Dordrecht, Kluwer.

Piek Vossen. 1999. *EuroWordNet general document*. EuroWordNet (LE2-4003, LE4-8328), part A, final document deliverable D032D033/2D014.

Webster's New International Dictionary of the English Language. 1913. Springfield, Mass.: G.&C. Merriam.

Hamza Zidoum. 2008. Towards the Construction of a Comprehensive Arabic WordNet. In *Proceedings of GWC 2008*, pages 531–544, Szeged, University of Szeged, Department of Informatics.