Automatism and User Interaction: Building a Hungarian WordNet

Gábor Prószéky*, Márton Miháltz†

*MorphoLogic
Késmárki u. 8, H-1118 Budapest, Hungary
proszeky@morphologic.hu

†Eötvös Lóránd University, Institute of Informatics
Budapest, Hungary
mmarcy@inf.elte.hu

Abstract

This paper attempts to provide an account of an ongoing project on developing methods with software implementation for building multilingual lexical databases based on Princeton WordNet. The objectives of this project are not unique; several similar projects have been carried out to different stages. We have been implementing a combination of manual and automatic techniques. The result is an effective procedure of building lexical nets with acceptable precision. As the project has been in progress for several months now, our account is according to the partial results we achieved so far.

1. Introduction

Our work so far has focused on the construction of the nominal part of Hungarian WordNet. We have started out from scratch, unlike many EuroWordNet participants with already existing lexical resources (Kunze et al., 1998). Our approach also differs from the general EuroWordNet approach involving the manual construction of Base Concept sets and continuing from there, adjusting to interlingual information (Vossen (ed.), 1999).

We employed the initial hypothesis that nominal hierarchies in English and Hungarian should be similar, at least for certain domains. This enabled us to formulate our task to attaching Hungarian nominal entries of a Hungarian-English bilingual dictionary to Princeton WordNet 1.6 synsets. This way, the English nominal hierarchy of WordNet serves as a skeleton structure to support the construction of the core Hungarian nominal WordNet. This approach was also taken up in the initial stage of the construction of the Spanish and Catalan WordNets (Farreres et al., 1998; Atserias et al., 1997), relying on a similar assumption. Our hypothesis also partly refers to previous work carried in the ACQUILEX projects, where a limited set of Spanish and Dutch nominal lexical entries were successfully linked automatically to a taxonomy extracted from LDOCE (Copestake et al., 1994). Furthermore, examining the Hungarian nominal taxonomies extracted from our monolingual dictionary, presented below, we have found that hierarchies for the concrete nominal domains (nouns denoting objects) seem to be similar to those in WordNet.

The resulting skeleton Hungarian WordNet structure, based on the English WordNet hierarchy is then analyzed for incorrect links, resulting from lexical gaps, incorrect information from the bilingual dictionaries and the errors of automatic methods. Attaching further nominal entries from a larger bilingual dictionary, a thesaurus and entries and definitions (serving as glosses) from a monolingual dictionary will enrich this core structure.

The remaining part of this paper is organized as follows: in the next section, we will give a review of the electronic resources we rely on for our work, with emphasis on information in support of our initial hypothesis. Section 3 gives an overview of the various automatic, semi-automatic and fully manual methods we are using, and how we integrate the information from the different sources. Finally, Section 4 comprises our conclusions and the directions of the further phases of our project.

2. Resources and Their Utilization

Besides WordNet 1.6, we have several electronic resources at our disposal: English-Hungarian bilingual dictionaries, a monolingual Hungarian explanatory dictionary, and a Hungarian Thesaurus.

2.1. Bilingual Dictionaries

MorphoLogic's English-Hungarian bilingual electronic dictionary contains entries for 17,801 Hungarian nouns with 12,440 English translations included in WordNet. The dictionary has been converted to a database of English-Hungarian word pairs with symmetrical translation relations (Prószéky et al., 2001). The entries of the Hungarian side constitute the basic set used for the various attachment procedures (Section 3.).

A much larger bilingual, a version of the renewed *English-Hungarian Academic Dictionary* (Országh–Magay, 2001) will be used for further improvement of the Hungarian WordNet structure. It contains over 150,000 Hungarian entries, with English translations covering more than 80% of WordNet's entries.

2.2. The *ÉKSz* Monolingual Dictionary

An electronic version of the Hungarian explanatory dictionary *Magyar Értelmező Kéziszótár* (ÉKSz) (Juhász et al., 1972) was converted to XML format. Figures for the nominal part of the ÉKSz monolingual dictionary are presented in Table 1.

Headwords	42,942
Definitions	64,146
Definitions annotated with usage codes	31,023
Headwords with translations in WordNet (through the smaller bilingual)	10,507
Monosemous entries	30,062
Average polysemy count (polysemous entries only)	2.65
Average definition length	5.22 words

Table 1: Figures for the ÉKSz monolingual

In order to aid the construction of the nominal Hungarian WordNet, information is acquired from the monolingual dictionary in several ways. First, programs were developed to parse each dictionary definition and extract semantic information. In 83% of all the definitions, genus words were identified, which can be accounted for as *hypernym* approximations of the corresponding headwords. In about 1,700 cases, the identified genus word was either a group noun, or a word denoting "part" relationship. For example, consider the ÉKSz entries for *alphabet* and *face:*

Alphabet: *The **set** of letters used for…*
Face: *The **part** of the head that…*

Using morpho-syntactic information, the meronym or holonym word (in our example: *letter, head*) could be identified instead of a genus word. This method provided *holonym/meronym* word approximations for 2.7% of all the headwords (only distinguishing between "part" and "member" subtypes of holonymy, as opposed to the 3 types represented in WordNet (Miller, 1990)). 13% of the definitions consisted only of a single noun. These function as *synonym* words for the corresponding sense of the headwords, which are mostly infrequent variants or compounds.

These simple methods provided us with hypernym, holonym and synonym words for 99.2% of all the senses of 98.9% of all the nominal dictionary entries. Such information extracted from machine-readable dictionaries can be used to build hierarchical lexical knowledge bases (Copestake, 1990), or semantic taxonomies (Rigau et al., 1998). The extracted genus word approximations can yield a hierarchical taxonomy of the nominal dictionary entries, organized by hypernym relations, providing a very versatile resource for the construction of our Hungarian nominal WordNet. However, in order to get hypernym relations between senses, the identified genus words have to be disambiguated, which means the hypernym sense must be selected from the senses corresponding to the genus word.

We are experimenting with several heuristics, relying on the work by Rigau et al. (1997) and Copestake (1990) to achieve an automated process of genus word disambiguation. About 70% of the genus terms are monosemous in the monolingual dictionary, in these cases the hyponym senses are attached to them directly. Another heuristic utilizes the *usage codes* available for about 30% of the candidate senses (see Table 1). Semantic codes, designating semantic domains such as S*port, Medicine, Science, Religion* etc. can be tested, if available, for compatibility between the hyponym and the candidate hypernym senses. The pragmatic codes, referring to typical language use, such as *technical, slang, vulgar, intimate,* etc. are also used: senses annotated slang, vulgar etc. are more unlikely to be used as genus terms. A third heuristic assigns the first sense occurring in an entry, relying on the fact that senses are ordered by usage frequency, and the most used senses are more likely to be used as hypernyms. A fourth heuristic tries to measure semantic similarity among definitions by means of determining the number of lemmas shared by both definitions. A fifth heuristic will rely on the conceptual distance formula, which measures semantic similarity between concepts using WordNet as a hierarchical knowledge base (Rigau et al., 1997). Application of the conceptual distance formula is discussed in more detail in Section 3.1.2.

Each heuristic will assign a score for the candidate senses, and the ones bearing the highest score will be linked to the hyponym senses. As work is still in progress for the disambiguation, it is early to report on the precision of the algorithm. Moreover, considering reports on previous works, it is likely that further manual and automatic assortment and/or verification of the resulting hierarchies will be necessary in order to attain a well-structured taxonomy (Rigau et al., 1998).

Some sample subsections of the resulting taxonomies were examined in order to investigate semantic similarities and differences between the parallel structures of the Hungarian hierarchy and WordNet. The most frequent difference originates from the fact that the hypernym trees in WordNet are quite detailed, often having 7-9 levels, while the Hungarian hierarchies tend to be more shallow, usually consisting of only 3-4 levels. The situation seems to be similar to previous projects constructing lexical hierarchies from machine readable dictionaries, for example in the Czech WordNet project (Pala & Sevecek, 1999). In the example in Figure 1, the two numbers following the Hungarian nouns refer to their homonymy and sense identifiers in the Hungarian taxonomy.

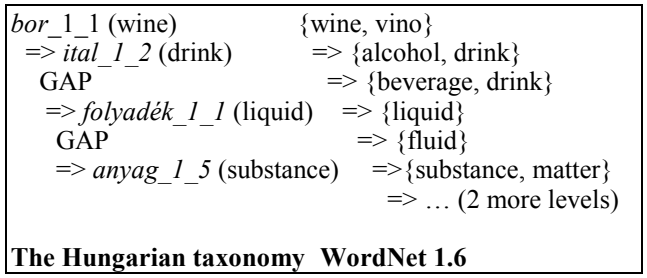

Figure 1. Lexical gaps in the Hungarian and English hierarchies

There are also cases where gaps occur in the WordNet hierarchy. For example, one would find the following classification for the sense *ló_1_1* (horse, as an animal) in the Hungarian taxonomy: *ló_1_1 => háziállat_1_1* (domestic animal) *=> állat_1_1* (animal). Whereas in WordNet, between the synsets *{horse, equus caballus}* and *{animal, animate being}* there are 7 levels in the hypernym tree, but none of them correspond to a "domestic animal" sense.

There are interesting cases resulting from the fact that the ÉKSz monolingual sometimes contains only holonym/meronym links for given entries. To cite an example: on the one hand, in the Hungarian taxonomy, the

sense *alany_1_1* (subject, of a sentence) is linked with a (part) holonym relationship directly to the sense *mondat_1_1* (sentence). On the other hand, in WordNet, {*subject*} is first classified as a kind of *{constituent, grammatical constituent},* which is then linked by (part) holonymy to *{sentence}*.

Based on the samples examined, besides the lexical gaps at both sides, the two hierarchies seem to differ most at the highest, most abstract levels, where the Hungarian taxonomies are very often unelaborated and confounding, or contain circular references. Nevertheless, we have not found evidence strongly contrasting our basic hypothesis, and our approach of attaching Hungarian nouns to the WordNet hierarchy seems maintainable for the initial stage of our work.

2.3. Other resources

We also have at our disposal a Hungarian electronic thesaurus. The *Magyar Szókincstár* contains 25,500 entries with synonyms and 14,400 entries with antonyms. Entries are linked to separate sets of synonyms for the various senses. Most of the synonym and antonym words are annotated with language usage labels.

Finally, we have a software interface developed by Dániel Nagy, (2001) for use in the manual linking procedure. The use of the Internet makes it possible for our contributing experts to work independently. Application of the manual tool is further discussed in Section 3.2.

3. The Methods

We are using three kinds of methodologies in order to achieve the task of linking Hungarian nouns to the WordNet synsets: automatic, manual and semi-automatic.

The automatic methods rely on the bilingual and monolingual dictionaries, and on the extracted semantic information, applying heuristics developed for the construction of the Spanish and Catalan WordNet (Farreres et al., 1998; Atserias et al., 1997). We chose to test these methods because the resources available to the Spanish and Catalan Research Group are closest to our available resources, considering the participants in the EuroWordNet project (Vossen et al., 1999).

The manual methods provide for a framework of top-down construction of the Hungarian nominal WordNet. This process will be supported by the use of the hierarchical taxonomies gained from the monolingual dictionary. This will enable the semi-automatic attachment of as much hyponym words as possible once a Hungarian word is linked to a WordNet sense.

The sets of candidate links produced by the different methods are inspected manually, and a common set is constructed. In the following section, the various methods and the means of integrating the results is discussed in detail.

3.1. Automatic methods

We are using two kinds of heuristics, described in Atserias et al. (1997) to automatically formulate candidate links for the Hungarian nouns to WordNet synsets. The first group of heuristics relies on information found in the bilingual dictionary and the structure of WordNet, while the second type relies on the information extracted from the monolingual dictionary.

3.1.1. Using the bilingual dictionary

Of the 17,800 Hungarian nouns forming the initial set, about 7,000 have translations in English which each belong to only one synset in WordNet. These nouns are classified into four groups, based on the nature of the Hungarian-English translation relationships (one-to-one, one-to-many, many-to-one or many-to-many). Then, for every noun in each class a hypothetical link is produced to the unique synset containing the translation(s). Atserias et al. report on different kinds of precision for the four classes, ranging from 85% to 92% correct connections (1997). Based on preliminary investigations, the average amount of correct links produced seems to be somewhat lower in our case. This is probably owing to the fact that the bilingual dictionary often either refers to senses not found in WordNet, or provides translations that correspond to hyponym senses of the Hungarian noun.

For the Hungarian nouns with polysemous translations in WordNet, the *Variant Criterion* and the 4 *Structural Methods* are being applied. These heuristics try to find common information between the English translations and WordNet. The *Intersection Criterion*, for example, will assign a Hungarian word to a synset if the synset is shared by at least two of the word's translations. In the Spanish experiments, precision is reported to be between 58% and 85% for these criteria (Atserias et al., 1997).

3.1.2. Using the monolingual dictionary

The ÉKSz explanatory dictionary contains *Latin* equivalents for about 1,600 nominal entries. These are mostly names of animal and plant species, taxonomic groups, diseases and chemical substances. Since WordNet 1.6 is very elaborate on Latin translations for such nouns, this provides for a reliable way for the linking of the Hungarian nouns. This method produced links for a small set of about 1,200 Hungarian nouns and corresponding definitions to WordNet, with the rate of correct connections estimated to be over 90%.

The second type of our automatic methods that utilize the monolingual dictionary relies on the extracted genus information (see Section 2.2). Following Atserias et al. (1997), we are applying the *Conceptual Distance formula* for the English translations of each headword-genus, or headword-holonym word pair we identified in the dictionary. The Conceptual Distance formula, introduced by Agirre et al. (1994), selects those two closest concepts in WordNet which represent the two input words. In the case of headword-genus pairs, the hypernym structure of WordNet is used as a semantic network for the heuristic, while for the ÉKSz headwords with holonym/meronym word approximations, the structures determined by WordNet's holonym links, are used.

The results of the application of the Conceptual Distance formula not only produce candidate links for the input Hungarian words, but can also be used as a heuristic to disambiguate the Hungarian genus word in the monolingual dictionary, contributing to the construction of the Hungarian nominal taxonomy (Rigau et al., 1997).

3.2. Manual and semi-automatic methods

A set of Internet-based software tools has been developed for manual disambiguation of the Hungarian nominal entries against WordNet.

An example for the task formulated here is depicted in Figure 2. The two Hungarian words *ló, lovag* have two English translations in the Bilingual, which belong to several synsets in WordNet (three of them displayed here). In Figure 2, solid lines represent connections between Hungarian and English words provided by the bilingual dictionary, and connections between the English words and WordNet senses, provided by WordNet. The user's disambiguation task consists of finding the correct subset of the edges connecting the Hungarian Words to the WordNet senses through the translations. Correct links are marked by dashed lines, incorrect ones by dotted lines.

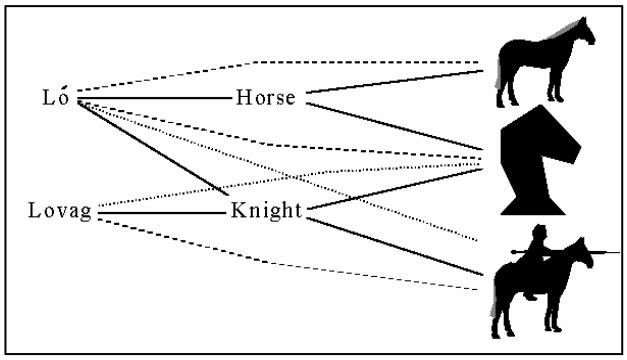

Figure 2. Relation between Hungarian words, English words and WordNet concepts

For human experts, the system offers a web page, over which the expert can answer questions provided by the central server maintaining the database. Experts are exposed to dialog boxes: if the word in question does mean the concept outlined below by English synonyms and a definition, then the human expert is supposed to press the Yes button. (Nagy 2001).

After the semantic taxonomy is extracted from the ÉKSz monolingual dictionary, it can be used in conjunction with the already available information gained from the previous steps and the English WordNet's structure to support the manual processing. The order of the manual disambiguation of Hungarian words with polysemous English translations will follow the top-down order (starting with abstract senses) of the English WordNet's hierarchy. Thus, once a Hungarian word is linked to a WordNet sense, hyponym words of its various senses can be disambiguated automatically against WordNet synsets, making use of the parallel structures of WordNet and the Hungarian taxonomy.

For example, let us suppose that the Hungarian word *állat* (`animal') has already been linked (either manually or automatically) to the WordNet synset {*animal, animate being, beast, brute, creature, fauna}*. *Állat* has 3 different senses in the Hungarian taxonomy, one of which has a hyponym pointer to (a sense of) the word *ló* ('horse'). The word *ló* has 3 English translations in the bilingual dictionary, which belong to 8 different synsets in WordNet . Which of those 8 synsets should *ló* be linked to? To answer the question, conceptual distance (see Section 3.1.2) is calculated between *animal, animate being,...* and the 8 candidate synsets. The candidate synset horse, *equus caballus* will show the smallest distance from the hypernym synset *animal, animate being, ...,* thus, Hungarian word *ló* (with the sense determined by the hypernym *állat*) can be linked to horse, *equus caballus*.

Some kind of a threshold condition is also built into the algorithm, which will prevent links to existing but incorrect WordNet senses (e.g. in cases where a Hungarian word has a hyponym sense that does not have an equivalent meaning in WordNet).

3.3. Putting it together

The result of the methods discussed above will be evaluated, based on random samples. Then all the possible intersections of the sets of results produced by the different methods will be evaluated, and only the results obtained by the combination which produces the highest accuracy will be considered, in order to ensure the precision of the core Hungarian WordNet structure (see Atserias et al., 1997).

4. Conclusions and Further Work

After the linking of the Hungarian entries of the bilingual dictionary to the WordNet semantic nodes is complete, further methods can be applied to enrich the resulting skeleton structure.

One way is with the aid of the *Magyar Szókincstár* thesaurus. With semantic disambiguation to decide which sense of a word the synonyms express, synonyms can be added to the Hungarian-English synsets. Antonyms to Hungarian words can be also added (antonymy is a lexical relation, therefore pre-existing WordNet antonyms cannot be used).

Daudé et al. (1999) describes a method for mapping multilingual hierarchies to WordNet using the relaxation labeling algorithm. Mapping the extracted Hungarian taxonomy to the Hungarian core structure using WordNet would provide the Hungarian WordNet with glosses, in addition to further synonymy and holonymy links.

In this paper we have described several methods we are using for the creation of the Hungarian nominal WordNet A combination automatic, manual, and semi-automatic are being used.. Automatic methods relying on the bilingual and monolingual dictionaries are used to link a basic set of Hungarian nouns to WordNet. The manual method relies on human experts, who are allowed to work independently. This process is supplemented by semi-automatic methods, which depend on taxonomies extracted from the monolingual dictionary. Our approach relies on the assumption that WordNet's semantic structure should provide us with an ample framework supporting the initial phase of our work.

5. References

Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarazza, and K. Sarasola, 1994. Conceptual Distance and Automatic Spelling Correction. In *Proceedings of the workshop on Computational Linguistics for Speech and Handwriting Recognition.*

Atserias, J., S. Climent, X. Farreres, G. Rigau and H. Rodríguez, 1997. Combining multiple methods for the automatic construction of multilingual WordNets. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark.

Copestake, A., T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A.

Samiotou, 1994. Acquisition of Lexical Translation Relations from MRDs. In *Journal of Machine Translation*, 3.

Copestake, A., 1990. An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. In *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*.

Daudé J., L. Padró and G. Rigau, 1999. Mapping Multilingual Hierarchies using Relaxation Labelling. In *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99)*.

Farreres, X., G. Rigau and H. Rodriguez, 1998. Using WordNet for building Wordnets. In*: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal.

Juhász, J., I. Szőke, G. O. Nagy, M. Kovalovszky (ed.), 1972. *Magyar Értelmező Kéziszótár*. Budapest: Akadémiai Kiadó.

Kunze, C., A. Wagner, D. Dutoit, L. Catherin, K. Pala, P. Sevecek, K. Vider, L. Paldre, H. Orav, H. Oim. 1998. First WordNets for BCs in French, German, Czech and Estonian. EuroWordNet (LE-8328) Deliverable 2D007.

Miller, G. A., 1990. Nouns in WordNet: a lexical inheritance system. In *International Journal of Lexicography* 3 (4), 1990: 245-264.

Nagy, D., 2001. *Computer Aided Methods for Lexical Database Compilation (Hungarian Nominal WordNet)*. Master's Thesis, Budapest University of Technology and Economics.

Országh, L., T. Magay 2001. *Angol-magyar nagyszótár*. Budapest: Akadémiai Kiadó.

Pala, K., and P. Sevecek, 1999. The Czech Wordnet. EuroWordNet (LE-8328) Deliverable 2D014

Prószéky, G., M. Mihăltz and D. Nagy, 2001. Toward a Hungarian WordNet. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 174–176.

Rigau, G., J. Atserias and and E. Agirre, 1997. Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. In: *Proceedings of the 35th Annual Meeting of the ACL*. Madrid, Spain.

Rigau, G., H. Rodriguez and E. Agirre, 1998. Building Accurate Semantic Taxonomies from Monolingual MRDs. In *Proceedings of COLING-ACL '98*. Montréal, Canada.

Vossen, P. (ed.), 1999. EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document.