Methods and Results of the Hungarian WordNet Project[1]

Márton Miháltz[1], Csaba Hatvani[2], Judit Kuti[3], György Szarvas[4], János Csirik[4], Gábor Prószéky[1], Tamás Váradi[3]

[1] MorphoLogic, Orbánhegyi út 5, H-1126 Budapest
{mihaltz, proszeky}@morphologic.hu
[2] University of Szeged, Department of Informatics, Árpád tér 2, H-6720, Szeged
hacso@inf.u-szeged.hu
[3] Research Institute of Linguistics, Hungarian Academy of Sciences, Benczúr utca 33, H-1068 Budapest
{kutij, varadi}@nytud.hu
[4] Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged, Aradi vértanúk tere 1, H-6720 Szeged
{szarvas, csirik}@inf.u-szeged.hu

Abstract. This paper presents a complete outline of the results of the Hungarian WordNet (HuWN) project: the construction process of the general vocabulary Hungarian WordNet ontology, its validation and evaluation, the construction of a domain ontology of financial terms built on top of the general ontology, and two practical applications demonstrating the utilization of the ontology.

1. Introduction

This paper presents a complete outline of the results of the Hungarian WordNet (HuWN) project: the construction process of the general vocabulary Hungarian WordNet ontology, its validation and evaluation, the construction of a domain ontology of financial terms built on top of the general ontology, and two practical applications demonstrating the utilization of the ontology.

The quantifiable results of the project may be summarized as follows. The Hungarian WordNet comprises of over 40.000 synsets, out of which 2.000 synsets form part of a business domain specific ontology. The proportion of the different parts-of-speech in the general ontology follows that observed in the Hungarian National Cor-

[1] The work presented was produced by the Research Institute for Linguistics of the Hungarian Academy of Sciences, the Department of Informatics, University of Szeged, and MorphoLogic in a 3-year project funded by the European Union ECOP program (GVOP-AKF-2004-3.1.1.)

pus and includes approximately 19.400 noun, 3.400 verb, 4.100 adjective and 1.100 adverb synsets.

In the following section, we describe our construction methodology in detail for the various parts of speech. In section 3., we present our validation and evaluation methodology, and in the last section we present the information extraction and the word sense disambiguation corpus building applications that make use of the ontology.

2. Ontology construction

The development of the HuWN followed the methodology that was called expand model by [8]. Although this general principle seemed applicable in the case of the nominal, adjectival and adverbial parts of our WordNet, naturally, some minor adjustments to the language-specific needs were allowed as well. In the case of verbs, however, some major modifications were necessary. Due to the typological difference between English and Hungarian some of the linguistic information that Hungarian verbs express through preverbs – related to aspect and aktionsart – called for an additional different representation method of Hungarian verbs than the one in PWN. This new representation, together with some other innovations in the adjectival part of the Hungarian WordNet are described in detail in [2] and a separate paper submitted to the Conference.

A second principle we decided to comply with was the so-called conceptual density, as defined by [6]. This means that if a nominal or verbal synset was selected for inclusion in the Hungarian ontology, all its ancestors were also added to the ontology. This way the resulting ontology is dense, in the sense that it does not contain contextual gaps. This fact has the advantage that later extensions of the HuWN can be performed by further extending the important parts of the hierarchies, without the need of constant validation and search for gaps in the upper levels.

During the construction of the HuWN there were several work steps when the usage of monolingual resources was necessary: the Concise Hungarian Explanatory Dictionary (Magyar értelmező kéziszótár, EKSZ) ([4]), a monolingual explanatory dictionary, the Hungarian National Corpus ([7]) and a subcategorisation frame table of the most frequent verbs in Hungarian, developed by the Research Institute for Linguistics of the Hungarian Academy of Sciences.

The relation types that have been retained from the Princeton WordNet are hypo- and hypernymy, antonymy, meronymy (substance, member and part), attribute (be_in_state), pertainym, similar (similar_to), entailment (subevent), cause (causes), also see (in the case of adjectives). Since the verbal relation indicating super- and subordination, called troponymy in PWN is called hypernymy in the version imported into the VisDic WordNet-building tool we have used, we have adopted the latter name. Some new relation types were also introduced, partly because of language specific phenomena – relations within the *nucleus-strucure* have to be mentioned here – and partly due to other, language-independent reasons – two new relations introduced in the adjectival HuWN, *scalar middle* and *partitions*, represent the latter

type of new relations. These are described in detail in a separate paper submitted to the Conference.

A primary concern when starting the ontology building was to provide a large overlap between the vocabulary covered by the Hungarian WordNet and other wordnets developed over the recent years. Accordingly, we have decided that we will take the BalkaNet Concept Set ([6]) (altogether 8,516 synsets) as a basis for the expand model, and find a Hungarian equivalent for all its synsets, or state if the given meaning is non-lexicalised in Hungarian.

2.1 Nouns

2.1.1 Translation of the BCS and adding the LBC

We first implemented the nominal part of the BalkaNet Concept Set (BCS sets 1, 2 and 3 together), consisting of 5,896 Princeton WordNet 2.0 noun synsets.

First, we applied several machine-translation heuristics, developed earlier ([3]) in order to get rough translations for as many literals as possible. This comprised about 50% of all BCS synsets. These were then manually examined, corrected and extended with further synonyms using the VisDic editor. We also allowed for many-to-one and one-to-many mappings between the ILI and HuWN synsets. The BCS synsets that remained untranslated by automatic means were translated manually and processed in a similar way. The lexicographers also linked related entries from the EKSZ dictionary to as many synsets as possible, and added definitions, based on EKSZ definitions.

As a starting point, we adopted all the semantic relations among the synsets from PWN 2.0. After the translation of all the BCS synsets to Hungarian was complete, we manually checked all the adopted relations and modified the hierarchies according to specifics of Hungarian lexical semantics.

Following the EuroWordNet methodology, we then added our Local Base Concepts (LBCs), synsets for basic-level and important Hungarian concepts not covered by the common core of the BCS. For this, we used a list of most frequent nouns in the Hungarian National Corpus and those used most frequently as genus terms in the definitions of the EKSz monolingual dictionary. For each of these, we identified the most frequent sense in the EKSz, then identified the subset for which no references were made in the Hungarian BCS. For these, we created 250 additional synsets, which constitute the local base concepts for Hungarian.

2.1.2 Concentric extension based on the ILI

After the creation of the concepts of the Base Concept Set and the Local Base Concepts, we decided to extend the Hungarian nominal WordNet concentrically, considering in several iterations the direct descendants of the ILI projection of the actual Hungarian WordNet as candidates. This way, the conceptual density criterion was automatically satisfied during the expansion, and we added general concepts from the upper levels of the concept hierarchy (since we started with the Base Concept Set).

Regarding the fact that upper level synsets usually have more than one hyponym descendants, in each iteration we had to select the 1-2 thousand most promising can-

didates from 30-40 thousand available. We used four, not necessarily concordant characteristics for ranking:

Translation: The concept candidate was preprocessable with automatic synset translation heuristics ([3]). This way the creation and correct insertion of the concept to the Hungarian hierarchy was easier to carry out, as one or more literals of the original English synset were available in Hungarian for the linguist expert.

Frequency: The concept had high frequency in English corpora (British National Corpus, American National Corpus First Release, SemCor). This usually indicates that the concept itself appears frequently in communication and thus adding it to the WordNet under construction was sensible.

Overlap with other languages: The candidate synset was conceptualized in WordNets for several languages besides English. This way we could maximize the overlap between Hungarian and foreign WordNets, that can be beneficial in multilingual applications like Machine Translation, and furthermore we could extend the ontology with such concepts that have been found useful by many other research groups as they added it to their own WordNet.

Number of relations: In the initial phases of the extension it made sense to take into account how many new synsets would become reachable by adding the one in question to the ontology. This way we could increase the number of candidates for later phases of the concentric extension.

2.1.3 Complete hierarchies for selected domains

As an additional extension method, we chose several domains for which all of the synsets in all of the hyponym subtrees in Princeton WordNet 2.0 were implemented in Hungarian. We did this to try to reach maximum encyclopedic coverage of the following areas:

- Geographic concepts and instances (countries, capitals and major cities, member states, geopolitical and other important regions, continents, names of important bodies of water, mountain peaks and islands)
- Human languages and language families
- Names of people
- Monetary units of the world.

We added 3,200 synsets based on these criteria

2.1.4 Domain synsets

In order to enable the coding of domain relations for synsets to be implemented in the future, we translated all the PWN 2.0 category and region domain synsets. We also extended the set of region domain synsets with a collection of specific Hungarian region names.

We decided to neglect the Princeton WordNet usage domain relationships because of several inconsistencies observed in PWN (e.g. in some cases, the usage classification pertains to all literals in a synset, while in other cases it doesn't.) Instead, we used a fixed list of our own usage codes, which could be applied individually to each literal using VisDic, providing a more flexible approach.

2.1.5 Proper names

National WordNets contain entity names among nominal synsets in a certain proportion. Among these are universal ones, like the world's countries, capitals, world famous artists, scientists or politicians, and ones that are important for that certain nation/country.

We added a considerable amount of the named entities that were found most useful for the Hungarian WordNet, after the following processing steps:

- Standardization (format and character encoding)
- Selection (selection of categories to incorporate to the ontology and selection of instances for chosen categories)
- Extension (we collected different transliterations, synonyms and paraphrases of the selected entities)

2.2 Verbs and adjectives

In the case of verbs, after an initial phase of applying the expand method, it became obvious that the simple translation of English synsets with the same hierarchical relations between them would not result in a coherent Hungarian semantic network, even if local modifications are allowed. Consequently, we decided to make more extensive use of our monolingual resources, and tried to apply a methodology that would both satisfy the need for alignment with the standard WordNet (at least concerning the core vocabulary) and the need for a representation that does justice to language-specific lexical characteristics of Hungarian.

Lacking frequency data of verb senses, we started out from the frequency data of Hungarian verbal subcategorisation frames, which in Hungarian have specific enough syntactic information to be close to determining sense frequency. We included all the senses of the 800 most frequent Hungarian verbal subcategorisation frames in the Hungarian WordNet and made sure they had English equivalents, but also allowing for approximate interlingual connections (*eq_near_synonym* relation). If the equivalent of a Hungarian synset was found outside of the range of the BCS, the criterion of conceptual density was followed in all cases.

In order to achieve a more consistent hierarchy of HuWN, we decided that although the Hungarian synsets themselves should be connected to the PWN equivalent synsets, their internal structure should be developed independent of the English one.

In the case of adjectives, the translation of the BCS synsets proved not to present such problems, and concerned only approx. 300 synsets. Given that these were all focal synsets of different descriptive adjective clusters, we followed the expansion method: we added the respective satellite synsets to the translated focal ones, and, if this was necessary, added the antonym half-cluster as well. This work, however, included some minor adjustments, since the lexicalized antonym pairs and their satellite synsets are highly language-specific, which should be reflected in the ontology. Some more structural changes implemented in the adjectival WordNet concerned antonym clusters which were not centered around a bipolar scale, but which had three circular antonym relations ([1]).

2.3 Adverbs

Considering the ratio of the parts of speech observed in corpora, we decided to add about 1,000 adverbial synsets in addition to the synsets of the localized BCS, that did not contain any adverb synsets.

Because of the lack of adverbial sense frequency data for Hungarian, we decided to translate about 1,000 most frequent adverbial senses in PWN 2.0. In order to accomplish this, we first selected PWN synsets containing at least one literal that occurred at least once in that sense in the SemCor sense-tagged corpus. Next, we added up all the frequencies of all the surface forms of all the adverbs in the American National Corpus for each PWN 2.0 adverb synset, and selected synsets with a score of at least 1. The intersection of these two sets formed 1,013 adverbial synsets, which were automatically and manually translated and edited as outlined above.

We then carried out a number of revisions in order to adjust for Hungarian semantics and morphology:

- Separated and added senses for adverbs that have both time and place meaning.
- For adverbs of place, we identified the possible direction subgroups determined by case suffixes, and made each subgroup complete.
- Merged PWN synsets that could be expressed by a single Hungarian adverb sense.

2.4 The financial domain ontology

Besides the construction of general purpose language ontologies, developing domain ontologies for specific terminologies is important, since the vocabularies of general language ontologies are rarely capable of covering the specific language of a special scientific or technical domain. The financial domain ontology connected to the general HuWN ontology served as a basis for information extraction application, described in section 0.

We used two different approaches to add domain-relevant terms to the Hungarian WordNet. First, we made use of the high coverage of Princeton WordNet. By manual inspection, we located 32 concepts in PWN that we found to contain relevant terms in the domains of economy, enterprise and commerce. We added the 1,200 synsets that are in the hyponym subtrees of these domain top concepts.

As a second step, we examined a domain corpus consisting of short business articles and collected candidate domain-relevant terms from the text. Those that were not already in HuWN have been added as synsets, along with their synonyms to the ontology. The following table summarizes the distribution of the domain terms as observed in a corpus, over the different parts of speech:

Table 1.

POS	Terms
Noun	2835
Adjective	270
Adverb	6
Verb	181
Overall	**3292**

3. Validation and Evaluation

3.1 Validation

In the final phase of the project, we focused on merging the parts of the ontology developed at the different project sites and performing several integrity and consistency checks, following [5]. The majority of the most frequent and serious problems were automatically identifiable with simple scripts and were then corrected manually. These included structural problems like:

- invalid sense ids
- same synsets connected with holonym and hypernym relations
- same synsets connected with similar to and near antonym relations
- duplicate synset ids
- duplicated relation between two synsets
- invalid characters in a literal, definition or usage example (character encoding issues)
- invalid relation types (mostly typos)
- improper linking to the EKSZ monolingual explanatory dictionary
- lexicalized (non-named entity) synset with empty or missing definition/usage example
- mismatching part-of-speech tag and id suffix
- Hungarian local synset with missing external relation
- direct circles in hierarchical relations
- duplicate literals in synsets
- invalid relation (connected synset does not exist or has different POS than required)
- the same definition is used in more than one synsets

We also checked some semantic inconsistencies that required manual inspection of the database by linguist experts, without major computer assistance:

- central synsets of two adjective clusters connected with near antonym relation (we considered these as improper uses of near antonym relation and hanged these to also see relations)

- not reasonable sense distinctions: two synsets could be merged as they represented practically the same concept (here we collected synsets that shared several literals in common.)

3.2 Evaluation method

In order to assess the relevance of synsets added to the Hungarian WordNet, we evaluated random samples from the whole WordNet, from the Base Concept Sets and from the whole hyponym trees we incorporated to the Hungarian Ontology, and compared them to the synsets that received the highest rank during one of the concentric extension phases.
The evaluation was performed in the following way:

1. We generated a random sample of 200 synsets from the concepts we wanted to evaluate.
2. Two native Hungarian speakers independently evaluated the importance of synsets according to their usefulness in a linguistic ontology. They had to assign a score ranging from 1 to 10 to each concept. The higher value they assigned to the concept, the more relevant it was in their point of view. The agreement rate of the annotators leveraged to all the samples was 78.67% (considering the agreement to be 100% in case they assigned the same value to the synset in question and 0% if the difference between their scores was maximal).
3. We took the average of the scores assigned by the two linguists for each synset and then calculated the average and deviance of scores over the 200 element samples.

3.3 Results

The columns of the following two tables represent the segments of the ontology from which we generated the 200 synsets large samples. These were:

NONBCS: the set of English synsets that are not among the base concept sets.
BCS1: 1st Base Concept Set
BCS2: 2nd Base Concept Set
BCS3: 3rd Base Concept Set
CONC_1: a random sample of synsets added during the first concentric extension phase
TREE: a random sample of synsets that were added during the extension of Hungarian wordnet by whole hyponym subtrees
CONC_2_CAND: a random sample of the candidates for the second concentric extension phase
LIT_FREQ: top ranked synsets from the candidates for the second extension phase using frequency-based ranking

ILI_OVL: top ranked synsets from the candidates for the second extension phase according to the number of foreign wordnets they appear in *Table 3.*

Table 2.

	NONBCS	BCS1	BCS2	BCS3	CONC_1	TREE
Mean	4.51	6.56	6.21	5.03	5.71	4.21
Deviance	2.48	2.78	2.20	2.45	1.71	2.61

.

Table 3.

	CONC_2_CAND	LIT_FREQ	ILI_OVL
Mean	4,25	5,26	8,32
Deviance	2,27	1,74	1,25

As a summary we conclude that it is worthy to construct evaluation heuristics for the selection of synset candidates to extend WordNets with. Some heuristics clearly helped to incorporate more useful concepts to the ontology than adding synsets without considering their relevance.

4. Applications

4.1 Information extraction

Our information extraction engine was developed to identify the event type (such as sales, privatisation, litigation etc.) and the participating entities (eg. the seller, buyer and the price in a sale) expressed in short business news texts.

We created so-called event frame descriptions manually after analyzing our collected business news corpus. Each frame description defines an event, and contains participants in specific roles that correspond to the main verb and its typical arguments. In the implementation of the IE engine, a parser first identifies the main syntactic constituents in the input text, and then it tries to match these to the elements of the candidate event frames. There are several kinds of constraints that have to be satisfied for a match. Lexical constraints can either be specified as strings, or as synset ids corresponding to hyponym subtrees of the HuWN ontology. Semantic constraints are expressed by so-called semantic meta-features, or basic semantic categories, such as “human”, “company”, “currency” etc. that are mapped to HuWN synsets and all their hyponyms. There are also syntactic and morphologic constraints, which

are checked against the output of the parser and the underlying morphologic analyzer. Finally, the IE engine ranks the candidate event frame matches for the output according to the ratio of event participants matched.

In this approach, the use of ontological categories allows for a simpler and easier to understand layout of the event frames. The main advantage of the use of synset ids and semantic types (as opposed to bare lexical listings) lies in the fact that the vocabulary of the IE engine can be easily customized and extended by adding new concepts to the ontology, without the need to modify the original event frame descriptions

4.2 Creating an annotated corpus for WSD

In parallel with the construction of the ontology itself, we selected 39 words that had several commonly used senses and built a lexical sample word sense disambiguation corpus for Hungarian. This corpus is freely available for research and teaching purposes[2] and consists of 350-500 labeled examples for each polysemous lexical item. The sense tags were taken from the synset ids of the senses of the polysemous words in HuWN.

The corpus follows the SensEval lexical sample format in order to ease its use for testing systems developed for other previous lexical sample datasets. The annotation was performed by two independent annotators. The initial annotation had an average inter-annotator agreement rate of 84.78%. Disagreements were later on resolved by consensus of the two annotators and a third independent linguist. The most common sense covers 66.12% of the instances and an average of 4 further senses share the remaining percentage of the labeled examples.

References

1. Gyarmati, Á., A. Almási, D. Szauter: A melléknevek beillesztése a Magyar WordNetbe. [Inclusion of Adjectives into the Hungarian WordNet] In: Alexin Z., Csendes D. (ed.): MSZNY2006 - IV. Magyar Számítógépes Nyelvészeti Konferencia, SZTE, Szeged, p. 117-126. (2006)
2. Kuti, J., K. Varasdi, J. Cziczelszki, Á. Gyarmati, A. Nagy, M. Tóth, P. Vajda: Hungarian WordNet and representation of verbal event structure. To appear in Acta Cybernetica (2008)
3. Miháltz, M., G. Prószéky.: Results and Evaluation of Hungarian Nominal WordNet v1.0. In Proceedings of the Second International WordNet Conference (GWC 2004), Brno, Czech Republic, pp. 175-180 (2004)
4. Pusztai, F. (ed.): Magyar értelmező kéziszótár. Budapest, Akadémiai Kiadó (1972)
5. Smrz, P.: Quality Control and Checking for Wordnets Development: A Case Study of BalkaNet. In Romanian Journal of Information Science and Technology Special Issue, volume 7, No. 1-2 (2004)

[2] Please contact the authors for information about obtaining the WSD corpus

6. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal of Information Science and Technology Special Issue, vol. 7, no. 1-2 (2004)
7. Váradi, T.: The Hungarian National Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas, pp 385-389 (2002)
8. Vossen, P. (ed.): EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014, (1999).