The Hungarian wordnet (HuWN)

University of Szeged, Department of Informatics
MorphoLogic Ltd.
HAS Research Institute for Linguistics

Contents

1. Introduction

1.1. Motivation

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 represent how linguistic knowledge is organized within the human mind (Miller et al. 1990). The first wordnet called the Princeton WordNet was created for English (Miller et al. 1990), which was followed by numerous wordnets all around the world. Wordnets for European languages have been developed mostly within the framework of the EuroWordNet and BalkaNet projects (Alonge et al. 1998, Tufiş 2004) among others.

Wordnets can differ in size, but they – especially the Princeton WordNet – are usually considered to be the largest database containing linguistic information for the given language. Thus, they can be used in various applications within the field of computational linguistics: word sense disambiguation, machine-assisted translation, document clustering, and so on.

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 the framework of the **Economic Competitiveness Operative Program (GVOP) 3.1.1-2004-05-0191** project (Alexin et al. 2006, Miháltz et al. 2008). As a result, HuWN now consists of over 40.000 synsets, out of which 2.000 synsets form part of a subontology in the business domain and later, 650 synsets were added from the legal domain.

The Princeton WordNet 2.0 served as a basis for the construction of HuWN; that is, synsets belonging to the BalkaNet Concept Set were selected from PWN 2.0 and then translated into Hungarian. These were then 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 not) was influenced by several factors such as the frequency of the concept or its presence in other WordNets (Miháltz et al. 2008).

Besides the construction of general purpose language ontologies, developing domain ontologies for specific terminologies is essential since the vocabularies of general language ontologies are rarely capable of covering the specific language terminology of a special scientific or technical domain. For this reason, two subontologies of the Hungarian WordNet were created, namely, an economic and a legal one.

1.2. Project data

Tender: GVOP-AKF-2004-3.1.1

Duration: February 01 2005. – April 30. 2007.

Participants:

MorphoLogic Ltd.
HAS Research Institute for Linguistics
University of Szeged

The objective of the project was to develop a 40 thousand-synset Hungarian wordnet, the source of which was PWN 2.0., the latest version of Princeton WordNet (PWN) at that time. This 40 thousand synset wordnet was realized in several steps. We first aimed to translate the 8.516 BCS synsets into Hungarian. Work-phases took place in the following iterative order:

1. To select synsets for the next work-phase from PWN synsets systematically, or on the basis of word and word-sense frequency data of the Hungarian language. The first is called the

expand model, the second *merge model*, since, in the later case, concepts are integrated in the ILI system subsequently.

2. Afterwards, to translate the new synsets into Hungarian, that is, to enter literals (word forms) into the synset; to draft a Hungarian definition; to compile a usage to illustrate the particular sense; and to add relevant references of The Concise Dictionary of the Hungarian Language (henceforth ÉKSz.) (if there is/are any).

3. Finally, to check synset relations.

Joining the EuroWordNet project has brought about considerable long-term advantages for R+D in Hungarian language technology, since the system offers an elaborate contact surface with semantic networks in various languages. EuroWordNet, as an intellectual product, unites – in an integrated way and with multilinguality in view – the advantageous features and theoretical results of independent research in the field of computational ontology of the past decades. The formalism of EuroWordNet provided a high-standard, cost-effective starting point for the realization of a Hungarian ontology.

Basic research had already showed up achievements in this field (Prószéky et al. 2001, Miháltz 2003) and the results made public at international scientific forums (Construction of the Hungarian nominal wordnet; some 10.000 Hungarian nouns linked to synsets in the Princeton WordNet) had been achieved part unaided, part with a modest amount invested.

1.3. The project objective, target groups

The main objective of the project was to create a large, highly-structured natural language concept set (ontology), the implementation of which has provided solutions for a number of scientific and technological problems.

Scientific objectives:

- To research and develop computer algorithms for the automated, heuristics-based support of ontology building, with the help of which manual work can be reduced to the control-integration phase.
- To examine to what extent concepts in Hungarian can be correlated with EuroWordNet Common Base Concepts and to what extent it is necessary to create – independently of EuroWordNet – a Hungarian top concept set.
- To examine the semantic description of the four parts-of-speech (noun, verb, adjective, adverb) of the Hungarian language in the WordNet formalism; to establish the necessary language specifics for Hungarian.
- To compare the taxonomy of Hungarian verbs with that of English verbs and to describe the differences and the Hungarian specifics.
- To describe frame information for Hungarian verbs.
- To isolate Hungarian business terminology, to examine the ways in which it is organized into an ontology and to look into the differences between the Hungarian and English business terminology.
- To research into fields of semantic analysis of electronic texts, such as word sense disambiguation, anaphora resolution and information extraction.

Technological objectives:

- To create a large, computational, natural language database (ontology) following the EuroWordNet formalism.

- To develop a business ontology and to integrate it into the general ontology.
- To develop software tools to support manual ontology building.
- To develop the prototype of an ontology-based, information extraction software module for the short business news domain capable of demonstrating the advantages of the application of a net of concepts.
- To create a 200.000-word control corpus by manual annotation for validation and to develop the required validation surface.

Social objectives:

In addition to purely scientific applications, Hungarian WordNet can also be utilized in various fields of education as it offers a user-friendly surface and it may serve as a visual aid in grammar teaching. Its applicability in language teaching is guaranteed by its standardized links to other wordnets. For example, making clear distinctions between the lexical differences of the learner's mother tongue and the target language may greatly promote the learner's acquisition of the lexical material of the foreign language. Hungarian WordNet can be utilized in developing "intelligent dictionaries", which make "getting the desired target language concept" possible in an interactive way, while keeping the danger of "mistranslation" low.

1.4. Project contents and activities

The main objective of the project was to develop a general ontology for Hungarian, to do related linguistic research, and to develop a prototype of an IE-system capable of demonstrating the practicability of the database.

Preparations took place in work-phase one. Within the frame of a short study, we developed the building principles and the methodology for the Hungarian version of EuroWordNet and the techniques for handling possible differences. In the same work-phase, we created the background database for the information extraction system (the database was being supplemented with short daily news) and the necessary infrastructure (server, clients, access technique) was developed.

In work-phase two, the actual ontology building started. Parallelly, semantic event descriptions (semantic frames) were being formed relying on the PWN structure. The consortium had already implemented an event description technology, which, till then, could not be based on a structured net of concepts but only on elementary semantic attributes. This event description system has been reconfigured in such a way that it can make full use of the potentials provided by hyponymy, hypernymy, synonymy and other relations in the ontology. Furthermore, manually annotated short news were used to create a test-database – only relevant pieces of semantic information were tagged – which served as a basis for testing and validation in the intermediate and final phases of the project.

In work-phase three, the consortium created the Hungarian wordnet database, which is part of EuroWordNet database comprising over 20 languages at that time. In this phase, an automatic semantic parsing system was developed, capable of matching the recognized nominal structures (noun phrases) with a corresponding concept or concepts in the ontology. Moreover, the semantic parser is also responsible for matching a given piece of short news with ontology-based semantic frames (developed beforehand) and for verifying the matching.

In work-phase four, we carried out further research in language technology in order to develop techniques that make the recognition and handling of semantic relations possible on a more sophisticated level. One field is word sense disambiguation. When using the ontology database, it may well be that the nominal structure within the text can be matched with more than one concept. Then, the correct sense can be selected on the basis of the syntactic and semantic environment. To solve the problem, linguist experts set up disambiguation rules and learning algorithms were also applied. Another domain was the resolution of distant, inter-sentential semantic relations and

references (typically anaphoras). Hungarian and other languages have a large number of tools for “distant” objects and references to events mentioned previously. To find them and then to take them into consideration during semantic parsing was the subject of our research. In this phase, the consortium developed the prototype of a web service using an automatic semantic annotation (parsing) technology, which embraces all the former results and developments. The main objective of the system is to extract information from business news with the help of an ontology, word sense disambiguation and parsing capable of detecting and allowing for inter-sentential semantic relations.

2. The construction of the Hungarian WordNet

2.1. VisDic as an annotation tool

VisDis was intended to be a freely available software developed for editing wordnet ontologies (Horak, Smrz, 2004). In both EuroWordNet and BalkaNet projects, VisDic was employed as an editor software in the course of wordnet building just as in case of the development of the Hungarian wordnet. The revision and correction of the automatically developed net of concepts was done with its help. The original version of VisDic was adapted to the building of the Hungarian wordnet since new functions and links were integrated into the database (e.g. links to the entries of the Hungarian Concise Dictionary, non-lexicalized synsets etc.).

2.2. Editing the synsets

The process of editing the synsets in VisDic happens as follows:

First, to check literals: to make sure whether the literal represents the given concept or not.

Then, to delete the unnecessary elements or, if necessary, to enter new ones.

To make sure that no identical literals remain at the *parent* and *child* nodes, therefore the literal at the *child* node gets deleted – in so far as there is/are an/other literal/s there; if there is no other literal at the *child* node, the synset gets the *t non-lex* label. (See 2.3.2.)

Afterwards, to produce a definition and a usage:

Where it is possible, to adopt an ÉKSz definition. In other cases, to take a PWN definition on the basis of which to produce a Hungarian rendering. If neither of the above ways are possible, to make up a proper definition based on linguistic intuition.

Then, to add a usage illustrative of the given synset. Either – just as in the above case – to write it taking the English usage as a starting point, or – independently of it – to create a suitable Hungarian sentence. Irrespective of the number of literals, only to add one usage.

In the next step, to enter the links connected with the corresponding entries in the ÉKSz, if there are any.

Subsequently, to check the relations[1]and – if necessary – modify them.

2.3. The non-lex problem

When rendering English synsets into Hungarian we often encountered the problem that English synsets do not always have direct equivalents in Hungarian. Possible solutions to this problem are presented below.

2.3.1. Non-lexicalized (non-lex) synsets

Creating the HuWN database practically meant rendering the PWN synsets into Hungarian, that is, we had to find Hungarian equivalents for English synsets. However, overlapping between two languages can never be perfect: due to the differences in culture, traditions and living conditions languages have concepts, words that are characteristic of the given language alone. They can only have approximate equivalents and cannot be expressed, translated with one word. Some of these words belong to a given culture: typically they are words of a historic tradition, folklore and

[1] Five major types of relations have been taken over and applied in the case of nominal synsets: synonymy, antonymy, holonymy, hypernymy and domain.

names of plays and meals belong here. Now, as regards the English-Hungarian language pair, though there can be found verbatim equivalents in the other language for the expressions presented below, they, however, do not reflect the feelings and moods they evoke – that is, what comes to a native person‘s mind when he hears them.

Hungarian examples:

Szent Korona – Holy Crown (it does not explicitly refer to the symbol of the Hungarian Kingdom)

Luca széke – Luca’s chair (it does not reveal anything about the related popular belief)

Máglyarakás – stake (in Hungarian, it is a kind of confectionery)

English examples:

Borderer – határvidéki (it is used to refer to people living along the border between Scotland and England)

Anglia – England in Latin (in Hungarian no distinction can be made, since the Hungarian equivalent of England is *Anglia*)

Another group of words belonging here includes elements that simply have no equivalent in the given language – to put it simply, there are no words for them. Very often, certain umbrella terms belong in this category that can only be expressed in the other language by using a paraphrase or giving a list. Here are a couple of examples:

Learned profession (a comprehensive name for law, medicine and theology)

Cycling (for both riding bicycles and motorcycles)

In order to not have “holes” in the constructed tree, that is, in order for the English and Hungarian wordnets to overlap to the highest possible degree, we had to come up with solutions for the proper handling of these synsets. To mark that these synsets do not exist (on the word level) in the lexicon of the given language, that is, they have not become lexicalized, the *non-lex* label has been introduced. These synsets give the concept corresponding the English synset in the form of a paraphrase, but no definition, usage and ÉKSz links have been provided and at the same time, the *non-lex* label has been added.

Also in the case of elements belonging to the first group, we decided – since translation cannot give back the concept altogether – to apply the *non-lex* label to the synsets and to provide them with a short description in the literal slot.

2.3.2. Technical non-lexicalized (t non-lex) synsets

While translating the English wordnet, it happened sometimes that two English words in hypernym-hyponym relation had one Hungarian equivalent. A narrower sense of this word is subordinated to a wider one, that is, the two concepts are separate on the conceptual level only, on the lexical level, however, it is impossible to find two distinct words for them. This, then, would have the result that in Hungarian the word is its own hypernym. In these cases we have two options:

a) if in the hyponym or the hypernym synset of a word not only the given word but other literals also occur, then the given word is **deleted** and in this way the problem of hypernym-hyponym overlapping is evaded; *kocka* in the hyponym synset has been deleted:

ENG20-03030489-n: kocka (kocka_1_5: Hozzávetőlegesen kocka alakú tárgy.)
{cube:5}hypernym
ENG20-03075421-n: **kocka**, dobókocka (**kocka_1_2: Dobókocka**.; dobókocka_1_1: Társas- vagy szerencsejátékban használt, lapjain 1-6 ponttal jelölt kocka.)

{dice:1}hyponym

b) if the hyponym synset contains only one literal (which is the same as the only literal in the hypernym synset), then the *technical non-lex* label is applied; it is always the hyponym synset that gets this label; both synsets contain the *függöny* literal:

ENG20-03037017-n: függöny (függöny_1_1: Ablakot, ajtónyílást stb. eltakaró, helyiségeket elválasztó, fent rögzített, félrehúzható csipkeszerű textília.)
{curtain:1}; hypernym
ENG20-03128470-n: függöny (függöny_1_2: Színpadnak a színpadot előadás alatt, után és a felvonások között a nézők elől eltakaró, nehéz kelméből való tartozéka.)
{drop curtain:1}hyponym; gets the *t non-lex* label; literal in parentheses; sense is 0.

In the case of the adjectival part of the ontology too, the *technical non-lex* label has been employed: since its construction is based on antonym-pairs and the associated, synonymous "satellite" synsets (see 5.1.), 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. However, the rules of wordnet building require that the focal and the satellite synsets 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 *technical non-lex* label.

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}.

Synsets with *technical non-lex* label – in contrast with synsets with *non-lex* label – have definition, usage and, in most of the cases, ÉKSz links. The reason why this solution was chosen that these synsets are existing concepts in Hungarian language that can be expressed with words and it is only due to the structure of the wordnet, that is, due to technical reasons, that we were compelled to provide them with the *non-lexicalized* label.

3. Nouns

Hereinafter, we present the main features of the nominal part of the Hungarian WordNet, the methodology, solutions that were employed in ontology building and the results of a test intended to control the quality of our extension methodology.

3.1. Methodological principles

When building the Hungarian WordNet – our main objective was to enter concepts that represent top-level, general linguistic knowledge, to which, later, smaller, domain-specific concept sets (such as a business ontology built subsequently) can be linked.

By the conceptual density criterion (which – in terms of practice – is considered a significant principle) it is meant that all those concepts should be entered in HuWN that are hypernyms of a given concept, that is, that are more general than the given concept. The conceptual density criterion is met, if after every extension phase, the top concepts of the nominal net are produced on the basis of the English wordnet and the incidentally missing synsets are added afterwards.

3.2. Nominal synsets

Dictionaries are usually structured on the basis of word forms: words are alphabetically listed in the dictionary, and their meanings are given one after the other. However, the most innovative aspect of wordnets is that lexical information is organized in terms of meaning; that is, a synset (the basic unit of wordnets) contains words of the same part-of-speech which have approximately the same meaning. Thus, it is synonymy that functions as the essential principle in the construction of wordnets (Miller et al. 1990). An example of a synset is the following:

{bicycle:1, bike:2, wheel:6, cycle:6}

Literals forming one synset are numbered as a word can have several meanings and it is important to represent that a word is synonymous with other words in one given sense. Thus, *cycle* occurs in five other synsets, including:

{cycle:1, rhythm:3, round:2}
{Hertz:1, Hz:1, cycle per second:1, cycles/second:1, cps:1, cycle:4}
{cycle:5, oscillation:3}

Synsets are connected to each other by means of semantic and lexical relations, yielding a hierarchical network of concepts. *Semantic relations* hold between concepts. In other words, not the forms but their meanings are related. Such relations include hyponymy and meronymy. On the other hand, *lexical relations* connect different word forms. For instance, synonymy, antonymy and different morphological relations belong to this group (Miller et al. 1990). Next, we will focus on the basic relations of wordnets – we provide definitions and illustrate them using nominal synset examples.

Hypernymy has a crucial role in forming the conceptual hierarchy in wordnets. A concept is a hypernym of another concept if it is a more generic term and the latter can be seen as an instance of the former (i.e. the IS-A relation holds between them) (Miller et al. 1990). For example:

{substance:1, matter:1} is *hypernym* of {fluid:2}, which is *hypernym* of {gas:2}

{furniture:1, piece of furniture:1, article of furniture:1} is *hypernym* of {wardrobe:1, closet:3, press:6}

Based on this relation, synsets can be organized into a conceptual hierarchy represented by a tree. Hypernymy is a transitive relation; that is, a synset usually has one direct hypernym, and it may have several hypernyms on different levels of the hierarchy. For instance, the direct hypernym of {bicycle:1, bike:2,

wheel:6, cycle:6} is {wheeled vehicle:1}, but its indirect hypernyms include {container:1}, {artifact:1, artefact:1} and {entity:1}. On the other hand, {bicycle:1, bike:2, wheel:6, cycle:6} is a hypernym of {mountain bike:1, all-terrain bike:1, off-roader:1} and {bicycle-built-for-two:1, tandem bicycle:1, tandem:1}, among others. This is illustrated in the following figure:

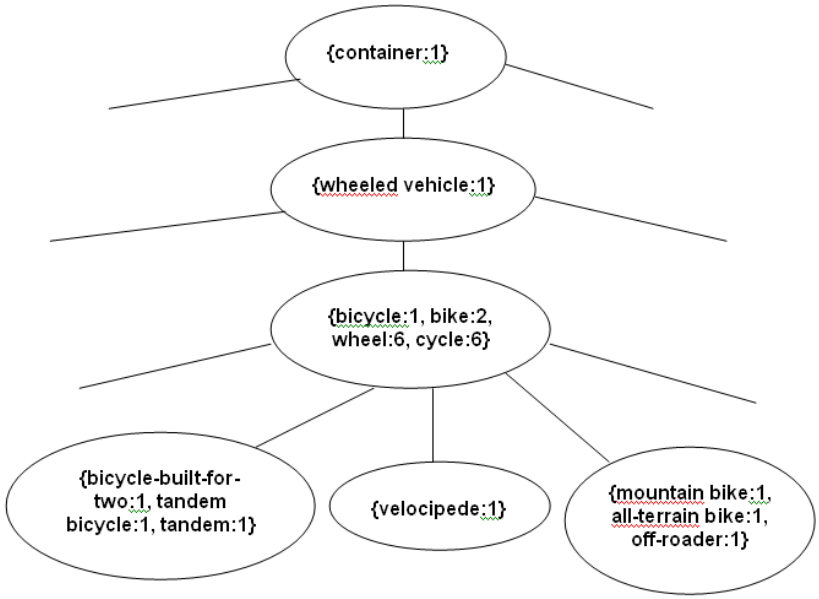

Fig. 1. Hypernyms and hyponyms of {bicycle:1, bike:2, wheel:6, cycle:6}

Holonymy and meronymy encode part-whole relations in wordnets. A concept is a meronym of another one if the former is a part of the latter (i.e. the HAS-A relation holds between them) (Miller et al. 1990). In the Princeton WordNet, holonymy is encoded by three different relations (Miller 1990), and in EuroWordNet there are two other relations besides these (Alonge et al. 1998). First, *holo_part* tells us that a thing is a component part of another thing:

{bicycle:1, bike:2, wheel:6, cycle:6}is *holo_part* of {pedal:2, treadle:1, foot pedal:1, foot lever:1}

Second, *holo_member* tells us that a thing or person is a member of a group:

{fleet:3} is *holo_member* of {ship:1}

Third, *holo_portion* refers to the *stuff* that a thing is made from (Miller 1990), but this relation links a whole and a portion of the whole in EuroWordNet (Alonge et al. 1998):

{joint:6, marijuana cigarette:1, reefer:1, stick:5, spliff:1} is *holo_portion* of {cannabis:2, marijuana:2, marihuana:2, ganja:2}
{bread:1} is *holo_portion* of {piece:8, slice:2}(EuroWN)

Fourth, *holo_madeof* encodes the *stuff* a thing is made from in EuroWordNet:

{paper:1} *has_holo_madeof* {book:2, volume:3}

Fifth, *holo_location* denotes a thing that is located within another place:

{oasis:1} *has_holo_location* {desert:1}

Holonymy and meronymy also allow us to visualize the relations between synsets as a tree structure. Here Figure 2 shows the parts of a bicycle (and the parts of a bicycle wheel):

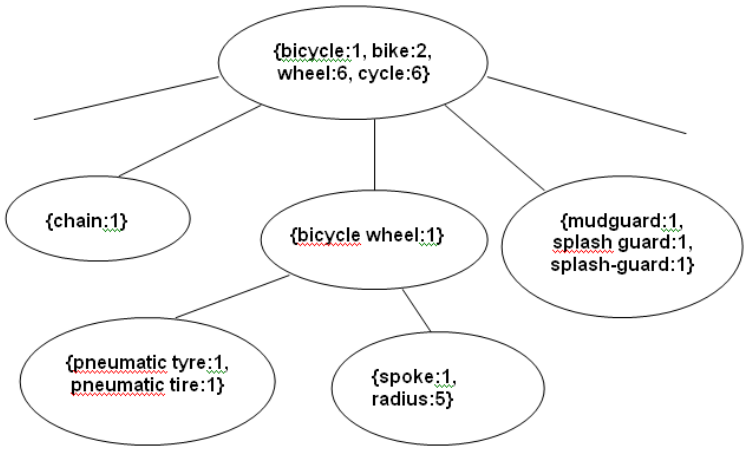

Fig. 2. Meronyms of {bicycle:1, bike:2, wheel:6, cycle:6}

Since a thing can function as a part of more than one thing – e.g. many vehicles have wheels –, it can have more than one holonym. This means that in a holonymic hierarchy, a leaf could belong to more than one tree. However, in this case it is more advisable to represent the hierarchy in a meronymic tree, where the top node is the part and the leaves of the tree are the entities that have the top node as a part of them. The following figure represents those entities that contain a handle as a part:

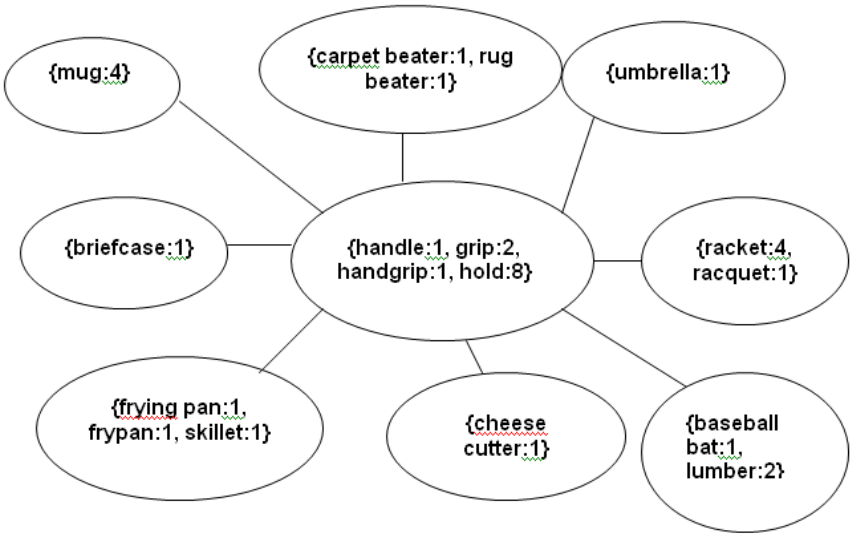

Fig. 3. Holonyms of {handle:1, grip:2, handgrip:1, hold:8}

3.3. Extension of the nominal net

In the preceding work-phases, we implemented the Hungarian representation of the synsets of the BalkaNet Concept Set (BCS), which is the common concept set of BalkaNet. The 8516 BCS synsets (5896 nouns) include concepts considered most important in the 8 languages of EuroWN and 5 other languages of BalkaNet, and reckoned basic in terms of ontology hierarchy. These concepts have been included in all the 13 languages, in this way guaranteeing a minimum level of interoperability among them. This nominal core ontology has been extended to 19.500 items, the process of which is presented as follows.

3.3.1. Local base-concepts

Following the EuroWordNet and BalkaNet methodology, we 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. The Hungarian nominal core-ontology is now quite likely to include – apart from the base-concepts of Balkanet/EuroWordNet – all the most important senses in the Hungarian language.

3.3.2. Concentric extension

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 extension, 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 candidates 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 (Miháltz 2003, Alexin 2006). 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.

In each phase, we chose the concepts ranked on the basis of frequency and overlap (and in phase one, on the number of relations) for the extension of the Hungarian ontology in such a way that we added 3-4 times as many synsets candidates with automatic translation as those without. In the so-called concentric extension phase, first, 2705, then 4385, finally 800 concepts have been completed.

3.3.3. Complete hierarchies for selected domains

In addition to the iterative, concentric, outward extension of the nominal stock of synsets, we selected some specific domains and translated every known PWN concept, that is, the whole hypernym subtree belonging to the given conceptual class was adopted. By doing so, we intended to reach maximum encyclopedic coverage for the given domains. This procedure was adapted for the following conceptual classes:

- geographic names (countries, capitals, major cities, (member) states within a country (e.g. US states), geographic areas (geopolitical regions), other regions, continents, names of important bodies of water (lakes, rivers, seas, bays, oceans, waterfalls), mountain peaks, islands;
- human languages (and language families);
- groups of people (nations, inhabitants of a region);
- monetary units of the world.

We have adopted 3,200 synsets based on these criteria.

By this method, 940 extra concepts have been added for the business ontology from the domains of economy, trade and finance.

3.4. Domain synsets

With the help of domain-relations introduced in PWN 2.0 we can represent relations that cannot be expressed by the usual semantic relations (in the case of nouns: hypernymy, holonymy, antonymy) and their role covers the function of the usage and domain labels of the conventional (explanatory) dictionaries. A relation represents a thematic/usage connection between a domain synset, as a comprehensive category and one or more domain term synsets, as elements. There are three types of domain relations: one expressing content/thematic/semantic relation (category); one expressing spatial relation (region); and one expressing a usage category (usage).

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 does not.) Instead, we used a fixed list of our own usage codes, which could be applied individually to each literal using VisDic (see 2.1. for details), providing a more flexible approach.

3.5. 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 in the following categories:

- geographic names (country, county, towns, other (mountain, river, etc.)
- names of establishments (companies, hospitals, theaters, cinemas, air companies, etc.)
- personal names (forenames, family names, names of famous people (artists, historic figures, etc.))
- titles (newspapers, books, novels, etc.)
- brand names (products, commodities)

Having had these lists the following processing steps were outlined:

- standardization (format and character encoding)
- selection (selection of categories to incorporate to the ontology and selection of instances for the chosen categories)

- extension (we collected different transliterations, synonyms and paraphrases of the selected entities)

3.5.1. NEs to be included as synsets

Selected elements of certain thematic lists have been directly adopted into HuWN. The categories are as follows:

- geographic names:
  - – countries
  - – Hungarian counties
  - – Hungarian towns
  - – cities of the world
- sights to see
- personal names:
  - – Hungarian forenames
  - – famous people

Every time an NE has a naturalized way of writing (a literal variant) in Hungarian, then the standard course to be followed is to represent the forms according to the Hungarian orthographical norms.

Subtasks concerning NEs to be included as synsets:

1. To manually select and label the NEs to be built in, to check and correct their written form.
2. During selection, to refine the bulk material, e.g. within the category of „famous people", to supply subcategories (painter, writer, poet, great military leader, politician, physicist, etc.)
3. To check whether the selected literals can be found in the English wordnet (there may be a problem with the automatic check if the Hungarian and English written forms differ or there is a special Hungarian name for an NE, e.g. Róma – Rome , Itália – Italy, Bécs – Vienna, Verne Gyula – Jules Verne). In this case, it is also feasible if before cross-checking them with the English wordnet, these synsets are automatically generated and when manual check takes place, they are linked to the English wordnet, if the English wordnet already has them.
4. To check whether the linking synset (hypernym) exists and, if necessary, to add it and translate it.

3.6. Evaluation of the extension methods

3.6.1. 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.

3.6.2. 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 1*.

	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 1: Ratings of samples

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

Table 2: Ratings of samples

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. Verbs

In this section, the construction of Hungarian verbal synsets will be presented along with specific problems and the solutions provided for them.

4.1. Basis methodological questions

After seeing some serious problems with the generally accepted *expand model* used when building wordnets (partly due to the weaknesses of the PWN and partly because of the differences between Hungarian and English), but being unable to merely rely on monolingual resources, we decided to try and find a compromise: using as many Hungarian resources, as possible, and keeping the general consortium principle of maximally aligning HuWN with PWN.

We took as starting ground all the subcategorisation frames of the most frequent Hungarian verb-lemmas (371 subcategorisation frames of 28 lemmas), which we turned into synsets. We also decided to allow multiple inheritance and artificial nodes in the verbal HuWN.

Instead of simply translating each English synset into Hungarian, and thus arriving at a one-to-one pairing of HuWN and PWN synsets, we allowed one-to-many and many-to-one correspondences between HuWN and PWN. Similarly, we accepted that the meaning distinction of the ÉKSz. and of HuWN might not be equally set, and consequently it was allowed, if needed, to have more ÉKSz. entries linked to one synset, and vice versa (see Figure 4). Exact match between an ÉKSz. entry and a synset indicated by "=" sign, approximate match indicated by "~" sign.

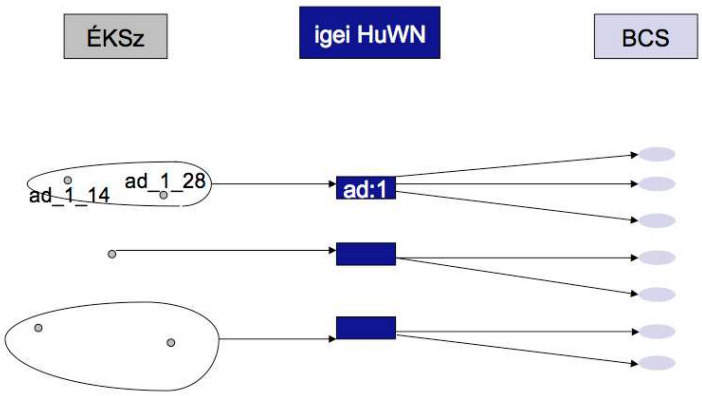

Fig. 4: Multiple linking relations

In accordance with the general consortium principle of maximally aligning HuWN with PWN, it was only required that each verbal HuWN synset have a clear indication of which PWN synset it is closest to, and in what way. These can either be:

- common synset ID (beginning with ENG20)
- ELR relation (*eq_xpos_synonym* or *eq_near_synonym*)
- hypernym synset that has an ENG20 ID, or a nucleus synset that is linked to a PWN synset

4.2. Relations

4.2.1. Intralingual relations:

SYNONYMY

- relation between concepts in one synset

- definition: mutual implication and co-temporality of the events[2] expressed by the verbs in question

HYPONYMY – HYPERNYMY

- Holds if
  - the hyponym would be considered a troponym according to Fellbaum (1998), i.e. the subordinate synset specifies a way of the event expressed by the superordinate synset (e.g.: *sleep – slumber*)
  - one of the arguments of the subordinate verb is the hyponym of the superordinate verb (*(living being) eat – (young animal or infant) suckle*)
- the subjects of the sub- and superordinate verbs have to refer to the same entity
- the following relations are inherited:
  - *subevent_nec_of* = the necessary subevent of an eventuality is also the necessary subevent of its subordinate

CAUSES

- Holds if
  - an event is the cause of another, i.e.: this latter event could not have happened without the former one (*overturn* (as in: *to overturn sg)* – overturn (as in: *sg overturns*)) (*strict causation*), or if
  - an event has triggered another one, but the latter one could not only have happened / taken place as a result of the former one (e.g.: *seat* (as in: *to seat sy.*) – *sit down*) (not strict causation)
- in both cases the direct object of the verb expressing the cause will be the subject of the verb expressing the consequence
- Theoretically, a clear distinction should be made between the two cases of causation as explained above, e.g. by introducing two distinct relation types. However, in the course of the project this modification was not introduced.
- Not to be mixed up with the relation called *has_consequence*!

NEAR_ANTONYM

- Holds if two synsets are each other's antonyms in any sense. Antonymy is in fact a cover term for several types of relations, which are not further detailed in PWN (see e.g. Vincze et al. 2008). We automatically took over this cover term in HuWN, and it was only in a later phase that we additionally introduced the *converse* relation (see 4.2.2.). We did not keep the automatically "inherited" *near_antonym* relation in cases when it was clearly unfitting, and we added it in a few cases where it was clearly the only relation that could define a synset in the network.

also_see and verb_group relations

With no available exact description of these relations in PWN, we did not delete them (except in cases where they were obviously not adequate, whatever semantic relation they may encode), but did not consider them as relevant in HuWN. It is important to note, however, that when a new Hungarian synset was created instead of two or more automatically generated ones, and the two (or more) original synsets were indicated as TNLs (see below), the new synset (with a HuWN-ID) did not automatically "inherit" the relations of the "old"ones. These had to be added manually in all cases, and this work has not yet been completed during the course of the project. The missing relations still have to be added in a later phase.

4.2.2. New relations introduced during the work on HuWN

HAS_CONSEQUENCE

- Holds if an event is the necessary consequence of another one.
  (e.g.: *imprison – hold captive, realize – know*)

[2] Events here stand for eventualities in Bach's (1986) sense.

- The relation shows from the direction of the synset lexicalizing the event bringing forth the other one in the direction of the synset lexicalizing the consequence.
- The subjects of the synsets lexicalizing the two events refer to the same entity.

TEMPORAL_PRECONDITION
- Holds if an event is the *necessary* precondition of another and precedes the latter one temporarily (the two event can overlap partially, but the precondition-event has to have a time point which precedes the consequence-event, and this must not be true the other way round.)
- The relation shows from the direction of the synset lexicalizing the later event to the synset lexicalizing the preceding event.
- As a default, the *temporal_precondition* relation connects verbs whose subjects refer to the same entity.

  E.g.: *give birth* <TEMPORAL_PRECONDITION *expect* <TEMPORAL_PRECONDITION *conceive*

However, since the temporal precondition relation can also hold between events / verbs of which either another argument (not the subject) refers to the same entity (*execute – quarter* (the direct object)), or between events / verbs of which neither respective argument refers to the same entity (*impregnate* <TEMPORAL_PRECONDITION *give birth*), these cases should ideally be clearly distinguished from the default *temporal_precondition* relation. Since the further specification of this relation has not yet been introduced in HuWN, but the number of the temporal_precondition relation is low altogether, this could be carried out any time work is continued in the near future.
- The inverse of the relation is automatically generated as *is_temporal_precondition_of*

SUBEVENT_OF
- Holds if an event 'A' temporarily includes an event 'B', which necessarily co-occurs with 'A', but the same does not hold the other way round, i.e. 'A' can happen without 'B', too. (e.g.: *sleep* ('A') *– snore* ('B'))
- The relation shows from the direction of the subevent, i.e. in the above case shows from *snore* in the direction of *sleep*
- The inverse of the relation is automatically generated as *is_subevent_of*

SUBEVENT_NEC_OF
- Holds if two distinct events necessarily occur together, e.g. *buy - pay*
- The relation is showing from the direction of the subevent, i.e. in the above case shows from *pay* in the direction of *buy*
- The inverse of the relation is automatically generated as *is_subevent_nec_of*

Both the relation *subevent_of* and *subevent_nec_of* were introduced as substitution for the PWN relation "subevent" automatically inherited. The subevent relation was neither explicitly defined, nor used with sufficient attention in PWN, resulting in an unreliable result when transferred automatically into Hungarian. We did not delete the original relation, however, in the hope to be able to check which of our new relations would be suitable instead of the 'old' one, but did not have time during the project to finish this work. Accordingly, the number of the 'old' *subevent* relation is much higher among the verbs than that of our newly introduced relations. Nevertheless, this work should be accomplished if the possibility arises.

CONVERSE
- Holds if two verbs lexicalize the same event from a different point of view. The verbs in the relation have to have at least two arguments.
  - E.g.: the subject and one of the complements "swop places":

    X people fit in the stadion. – The stadion houses X people. (2 arg. → 2 arg.)

or

X bought a car from Y. – Y sold a car to X. (3 arg.→ 3 arg. – direct object unaltered)

or

* the direct object turns into the subject (patient thematic role), the subject "disappears"
X mentions Y. – Y gets mentioned. (lexicalised in Hung.) (2 → 1 arg.)

AKTIONSART

Introduced for the cases when a verb can be clearly characterised with the help of an "aktionsart" according to Kiefer (2006), e.g. *inchoative* aktionsart in the case of *start:1*. An artificial node is created for these cases, which seems to be functioning as a root synset, but is defined as one that does not form part of the hyponym-hypernym hierarchy, e.g. AKTIONSART KEZDET / AKTIONSART BEGINNING. The relation points from the natural language synset towards the artificial node. The inverse of the relation is automatically generated as <<-- aktionsart.

4.2.3. Artificial nodes in the verbal HuWN

Artificial nodes can be added to HuWN in the following cases:

- for the structuring of an unmanageable amount of co-hyponyms, if no lexicalized expression is at hand to structure some of these. E.g.: MOZOG:2 (MOVE:2). Artificial nodes are indicated by CAPITALIZING all the letters in the literal.

- for the indication of so-called nuclei – see 4.3.2.1.

4.3.2.1. Nuclei and relations within a nucleus

Nuclei, as means of structuring Hungarian verbs have been introduced in HuWN in order to allow for the representation of aspectual information when expressed morphologically, through a verb-prefix. The notion of a nucleus was introduced relying on Moens & Steedman (1988).

The central notion of Moens & Steedman is an idealized event-unit that comprises three parts: a preparatory phase, a culmination point / telos and a consequent state that might be represented as **<a, b, c>** Their distinction rely on Vendler's aspectual classes but further refining it.

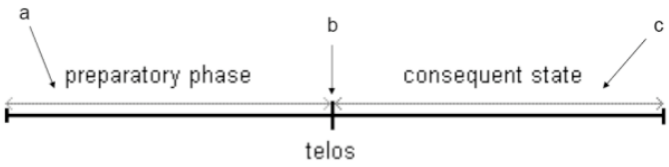

Fig. 5: Parts of an event

Moens & Steedman place this idealised event-unit beyond the level of linguistically manifested lexicalised meanings. The components of the event-nucleus are thus filled with meta-linguistic and not with lexicalised linguistic elements. There are linguistic tests with which one can test whether a lexicalized expression conceptualises one or more of the above nucleus-components. On the example of the eventuality lexicalised with the verbal phrase go out of the room: The existence of the first component can be tested by looking at whether the expression can be put into the progressive. An expression will be acceptable in the progressive if and only if the first component of its triad is conceptually present. The existence of the third component, which practically goes hand in hand with the presence of the second one, 12 can be tested by looking at whether the expression can be put into the perfective. Due to certain characteristics of the Hungarian language the easiest way we can test whether certain components of the triad are conceptualised is by translating the Hungarian sentence into English and putting the translated equivalent into Present Perfect / Progressive.

János éppen ment ki az épületből, amikor találkoztam vele.

János was going out of the building when I met him.

Mire Zsuzsa megérkezett, addigra János kiment az épületből.
By the time Sue arrived, John has gone out of the building.

As a result of the two tests we can see that the phrase go out of the building conceptualises all the three components of the triad:
<GOES TOWARD THE GATE, PASSES THE THRESHOLD, IS OUTSIDE>

Theoretically 2^3 different potential aspectual types may be distinguished according to the conceptual presence of the nucleus-components, listed as follows.

< □, □, □ > < □, b, c > < a, □, □ >
< a, b, c > < a, □, c > < □, b, □ >
< a, b, □ > < □, □, c >

The coherence of the nucleus components is more than mere temporal sequentiality, it is what Moens & Steedman call contingency — "a term related, but not identical to a notion like causality". The mutual dependency among the three components of the nucleus means that none of them can be seen as preparatory phase, culmination or consequent state per se. An eventuality that, based on the above tests, seems to possess a preparatory phase, but lacks both culmination and consequent state (could be marked as < a, □, □ >) cannot be seen as a preparatory process, as it does not precede anything. By analogy, an eventuality that, based on the above tests, seems to possess a consequent state but lacks a culmination (could be marked as < □, □, c >) cannot be seen as a consequent state, just like an eventuality with what seems to be a point of culmination, but lacking both preparatory phase and consequent state (could be marked as < □, b, □ >) cannot be interpreted as a telos. In other words, a triad having a consequent state implies that the triad also has a culmination point. However, the three respective components seemingly appearing on their own may easily be interpreted as corresponding to the notion *process* and *state* as used by Vendler and to the Bachian *point* expression.

Although the three non-complex eventualities (process, point, state) are not discussed further by Moens & Steedman, we deal with them in HuWN, and follow the above convention of showing the aspectual information in an ordered triple. Accordingly, the above listed possible combinations of the nucleus-components, each standing for one possible aspectual verb-subtype, are illustrated with examples, as follows:

<*ø, ø, ø*>	no example
<a, b, c>	*befelhősödik ('become cloudy')*
<a, *ø, c*>	no example
<*ø, b, c*>	*eltörik ('break')*
<a, b, *ø*>	no example
<a, *ø, ø*>	*fut ('run')*
<*ø, b, ø*>	*kattan ('click')*
<*ø, ø, c*>	*szeret ('love')*

Three of the possible combinations are excluded based on epistemologic grounds: (i) A nucleus having no components at all cannot be discussed neither conceptually nor linguistically. An eventuality (ii) having a preparatory phase and a culmination point, as well as one (iii) having a preparatory phase and a consequent state cannot be lexicalised due to the coherence of the telos and the consequent state. Besides the remaining five lexicalised possibilities of nucleus-component combinations we have, however, seen the need for marking a sixth possible aspectual type in HuWN. As mentioned above, in many cases linguistic tests in Hungarian are unreliable in the sense that they provide ambiguous results even for native speakers. For the sake of usability in Hungarian

language technology applications we considered it necessary to explicitly mark those cases in HuWN where the Hungarian test for the progressive did not result in a clearly grammatical sentence, but the English equivalent did. One such example can be seen in:

János éppen gyógyult meg, amikor huzatot kapott a füle és újra belázasodott.
John was getting better when his ear caught cold and he got fever again.

In cases like the above mentioned we decided to mark the first component of the nucleus ”unmarked”, designating this with an x: <x,b,c>

The notion of the nucleus in HuWN

Telicity in WordNet

As we have seen, the conceptual presence or absence of meta-language elements beyond the lexicalized expressions can be tested with the help of Moens & Steedman’s nucleus structure. The number of components a verb conceptualizes compared to an idealized complex event unit provides information on the telicity or atelicity of a given eventuality. If the third component of a nucleus denoted by a given verb is expressed, the eventuality is telic, if this component is not present, the eventuality is atelic.

We have decided to indicate telicity within the synsets on the level of literals, in the above manner:

(a,0,0) (0,b,0) (0,0,c) standing for processes, pointlike expressions and states.
(a,b,c) standing for a fully lexicalised nucleus-structure
(0,b,c) standing for an eventuality with telos and consequent state
(x,b,c) standing for the above mentioned case when the verb is underspecified as to whether the preparatory phase is conceptualised

Complex eventualities in HuWN

Besides the possibility of storing a minimal amount of aspectual information concerning the given literal in a verb synset, the relational structure of the wordnet and the nucleus taken as a single unit allow us to propose another extension to the verb synset structure. In the case of complex eventualities whose certain triad components are not only conceptually present, but are lexicalised, as well, the unity of these components can be represented. Although the structure of PWN is based on a hierarchical system, an alternative structure has already been accepted for adjectives in PWN. By analogy it must be possible to organise the verb synsets in a slightly modified way than nouns, as well. The tripartite structure described above may be mapped onto the system of wordnet in the form of relations. The metalanguage level described by Moens & Steedman’s nucleus structure can be mapped onto the level of lexicalised elements, represented by wordnet synsets. The connection of the two levels is shown below:

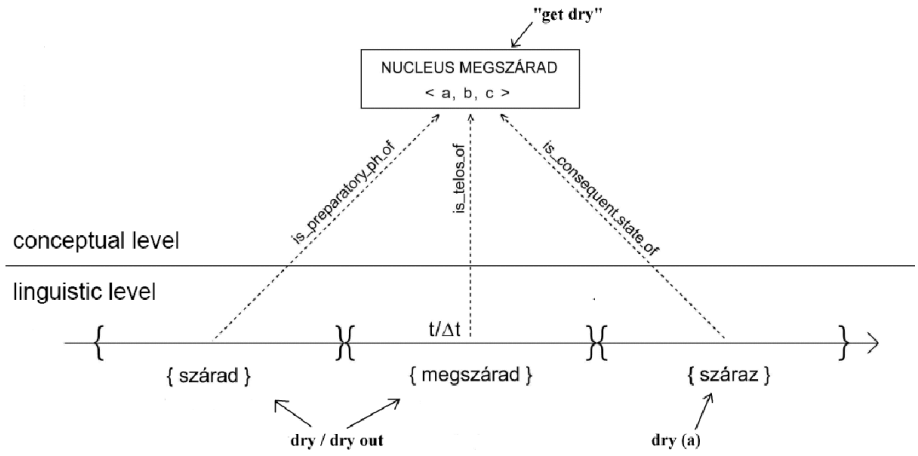

Fig. 6: Connection between the conceptual and linguistic levels

Artificial nodes introduced in HuWN are suitable for naming metalanguage nuclei, e.g. the complex eventuality denoting the change of state from wet to dry, in the above example.

The relational structure of the wordnet allows introducing three new relations according to the respective triad-components being related to the meta-language nucleus-unit, represented by an artificial node. These new relations point to the appropriate artificial node and they are called *is_preparatory_phase_of*, *is_telos_of* and *is_consequent_state_of*, respectively, based on the names of the different nucleus components.

Meanings that are lexicalized by a single verb in English but not in Hungarian can thus be distinguished: the same meaning might be present in Hungarian often as a verb with a preverb providing more aspectual information and as a verb without a preverb, more underspecified for aspectual information. In the above example, the Hungarian *szárad* and *megszárad* synsets are both equivalent to the English {dry:2}. Without integrating the nucleus system into the wordnet the synset *megszárad* could be placed into HuWN only as a hyponym of *szárad*, considering all the originally available relations. However, this kind of storage would not distinguish the different implicational relation between the above mentioned two meanings, but would merge them into a hyponym-hypernym relation. After having integrated the nucleus system into the wordnet, there is no need for an additional explicit relation between the components of a nucleus: they are already connected through the artificial node. Following the path of the relations *is preparatory phase of* and *is telos of*, it is easy to determine that the synset *szárad* represents the preparatory phase of the nucleus whose another lexicalized component is *megszárad*, hence *megszárad* implies *szárad*, while the implication does not hold in the other direction.

As we have seen, verbs belonging to the same triad (often with and without a preverb respectively) can be placed more accurately in HuWN with the help of the new relations. Furthermore, the relation is consequent state of is not restricted to verbs, the third component of the triad mentioned above is the adjective synset *száraz* ({dry:1}). This psycholinguistically relevant piece of information is present in HuWN but would be lost if we had strictly held onto the structure of PWN without the tools for representing triads.

The *causes* relation between nuclei

The *causes* relation, that exists in PWN typically between many synsets under two nodes: *change:1* and *change:2*, indicating the undergoing and causing of some kind of change, can be implemented between nuclei in the same two trees with less redundancy than without these artificial nodes. The meaning of the *causes* relation between nucleus nodes, is, following the figure below: a1 $\rightarrow$ a2, b1 $\rightarrow$ b2, b1 $\rightarrow$ c2, the arrow representing the *causes* relation.

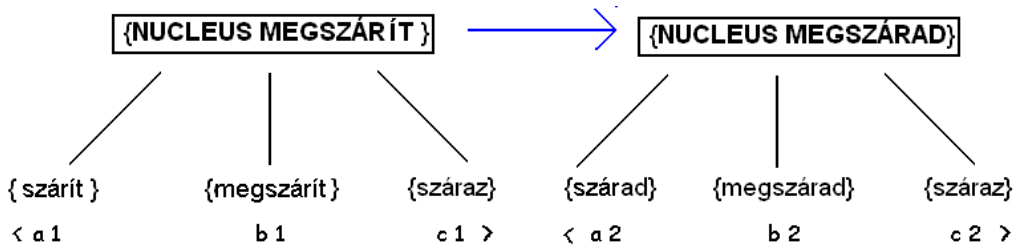

Fig. 7: The *causes* relation

4.2.3.2. Verbal non-lex and technical non-lex synsets

We decided to distinguish in HuWN between synset / concepts that are truly *not lexicalised* in Hungarian, i.e. representing a lexical gap (we marked these NL), and synsets that were automatically generated from PWN as Hungarian counterparts, but turned out to be superfluous, due to some reason. Typically this reason was that PWN had more synsets for apparently one and the same concepts, and we did not want to take over this synset-duplication. In this case, the two (or more) automatically generated synsets were marked TNL (technical non-lex), and a new node was created instead of these two (or more), which was linked to the PWN synsets it was the contraction of through an *eq_near_synonym* relation (indicated by the wavy line below).

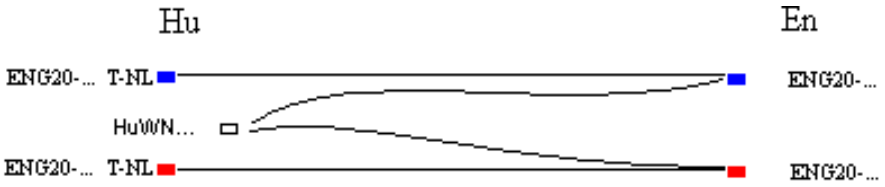

Fig. 8: The *eq_near_synonym* relation

Another similar case of using the TNL option is when there are two (or more) seemingly equivalent synsets in PWN (red and blue synset below, on the right hand side) but on a closer look it turns out that the one synset is fully elaborated (below, the red one), the other one is not. In cases like these, the automatically generated counterpart of the more acceptable synset is retained with an ENG20-ID (indicated below by the longer equation sign), while the counterpart of the less elaborated one is marked TNL and gets linked to the accepted HuWN synset by a new relation: *near_synonym* (represented below by the the green line, while its original hyponymy relation is deleted).

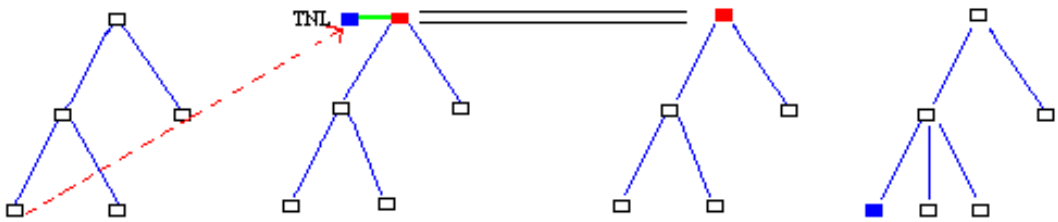

Fig. 8: The *near_synonym* relation

A third case of using the TNL option was when the same concept was represented by a different part of speech (POS) in Hungarian than in English. The automatically generated synset, of the same POS as in English, was marked TNL, and a new one was created, which got linked to the PWN synset by an *eq_xpos_synonym* relation. The *eq_xpos_synonym* and the *eq_near_synonym* relations are the only ELR relations used in HuWN.

4.3. Information on subcategorisation frames in HuWN

We indicated in each synset of the verbal HuWN the subcategorisation frame of its literals in a new XML tag called VFRAME. The subcategorisation frame was available in most cases in a table that had been put together manually in the Research Institute for Linguistics, and covers altogether some 17.000 subcategorisation frames (the VFRAME tag actually contains the identifier of this entry). The correspondence between the subcategorisation frame and the synset literal can be either exact or approximate (indicated by = and ~ respectively). Since some of the entries of the subcategorisation frame table were modified in retrospect, a semi-automatic checking of the available verb frame information in the synsets is still needed.

4.4. Initial steps of creating a new upper ontology

During work on the verbal WN we tried to deal with aspectual properties of Hungarian verbs. As a result, the need arose to include this aspect when dealing with the root synsets in the hierarchy, and to add some artificial nodes to the upper ontology, which complement the existing upper nodes. The current state of our proposed upper ontology is depicted in the picture below. This state is by no means to be considered final – further development should follow, if possible.

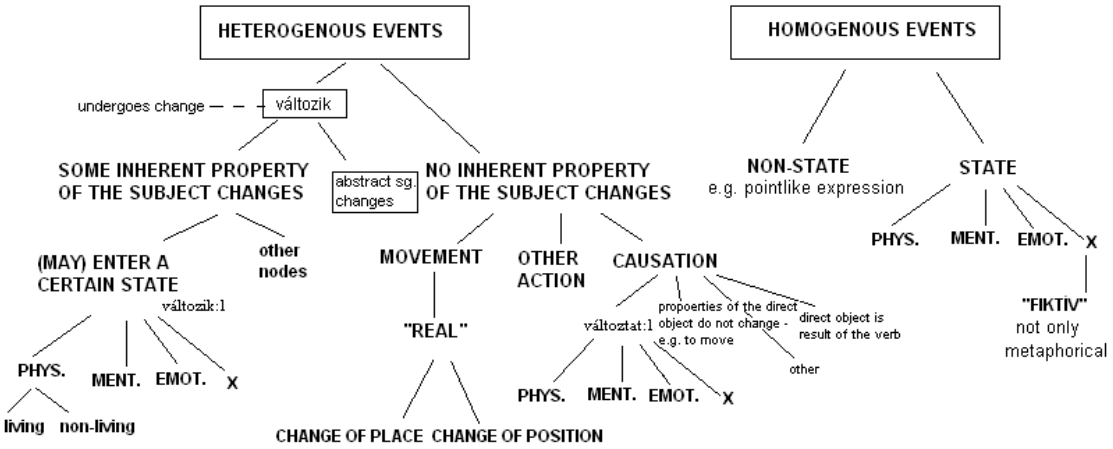

Fig. 10: Upper ontology of verbs

5. Adjectives

Basic – i.e. generally used – adjectival relations and new relations introduced in the Hungarian WordNet are presented here.

5.1. Adjectival relations

1. *near_antonym*

In the lexical database of *wordnet*, the position of words is determined by semantic relations. The structure of the adjectival wordnet radically differs from nominal and verbal structures. In the case of adjectives, the most substantial relation of the adjectival structure is *antonymy* as opposed to *hypo-hypernymy* in the case of nouns and verbs. As a result of this, the majority of adjectives forms a so-called *cluster structure*. Central synsets are those with an antonym-pair.

Example: {long:1}-{short:2}/ relation type: *near_antonym*

2. *similar_to*

Related adjectives of the given dimension are grouped around antonym pairs. These adjectives have no antonym pair, that is, they have no direct antonym. These so-called satellite synsets are linked to the focal synset with similar sense with *similar_to* relation. In this way, satellite synsets – through their focal synsets – will have an indirect antonym.

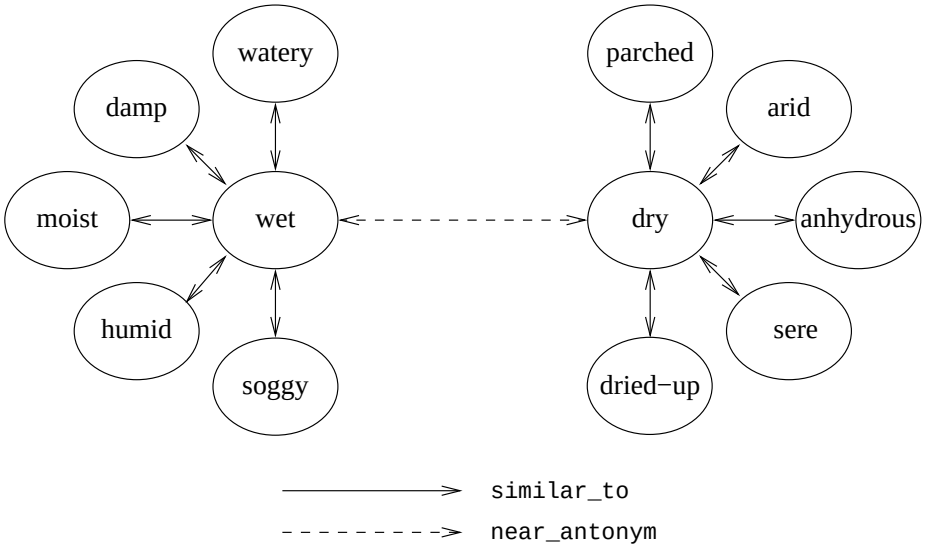

Fig. 11: Adjectival clusters

3. *also-see*

Also_see relation establishes connection between focal synsets with similar sense.

Example: {bad:1}- {evil:1; wicked:4} / relation type : *also_see*

{bad:1} is related to {good:1} with *near_antonim*. {evil:1; wicked:4} is the antonym of {good:3} and {good:1} and {good:3} are related with *also_see*.

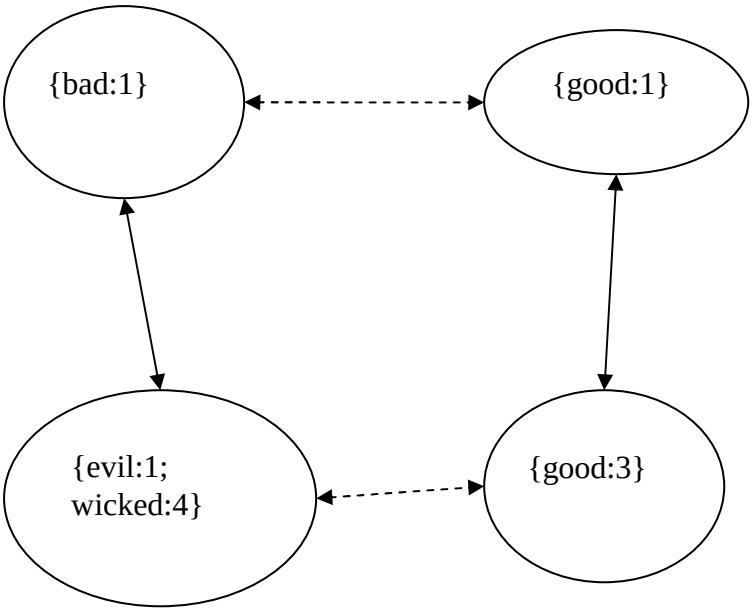

Fig. 12: The *also_see* relation

4. *be_in_state*

This type of relation links the adjectival synset to a noun, the adjective describes being in the state denoted by the noun.

Example: {evil:3; evilness:1} is characteristic of those who are {evil:1; wicked:4}.

5. *middle*

Not all adjectives can be described in terms of antonymy, because there are cases when between the two antonymous adjectives – at half-way – there is another adjective that marks out the center of the scale determined by the two polar, antonymous adjectives. This central adjective is not an antonym of the two polar adjectives, thus the *middle* relation has been introduced for its representation pointing to the polar adjectives from the center.

Example: {amphoteric:1, amphiprotic:1} is in *middle* relation with {acidic:1} and {alkaline:1, alkalic:1}in the Hungarian wordnet.

6. *partitions*

This relation links adjectival synsets to nominal ones where the scope of the adjective is limited and can only refer to a noun belonging to a given semantic class.

Example: {extinct:2, inactive:4} *partitions* {volcano:2} ;{ dormant:1, inactive:5} *partitions* {volcano:2}; {active:12} *partitions* {volcano:2} in the Hungarian wordnet.

When creating the adjectival part of the Hungarian wordnet, a number of different aspects had to be considered. Due to its character, it was created on the model of PWN and it strives to retain the PWN structure within the frame of the Hungarian language in the case of literals and relations as well. However, there are considerable lexical and association differences between the

two languages, and as a result of this, by merely translating the adjectival part of PWN we do not get the corresponding part in HuWN. studying the following, we give an account of the general and language-specific problems that occurred when building the adjectival part of HuWN. This, then, necessitated an investigation into POS-categorization and the introduction of new relations into the wordnet.

5.2. Adjectives in HuWN

However, the adjectival part of HuWn is not simply a translated version of PWN since its development needed thorough preparation and work. This proved to be necessary due to the differences between the English and Hungarian lexicon and word association on the one hand and we intended to eliminate the inconsistencies occurring in PWN.

5.2.1 Language-specific features

As it is known, antonymy can hold between words and not concepts, thus, it is not surprising that the structure of HuWN exhibits some minor differences compared to the one of PWN. The following figure illustrates such a case:

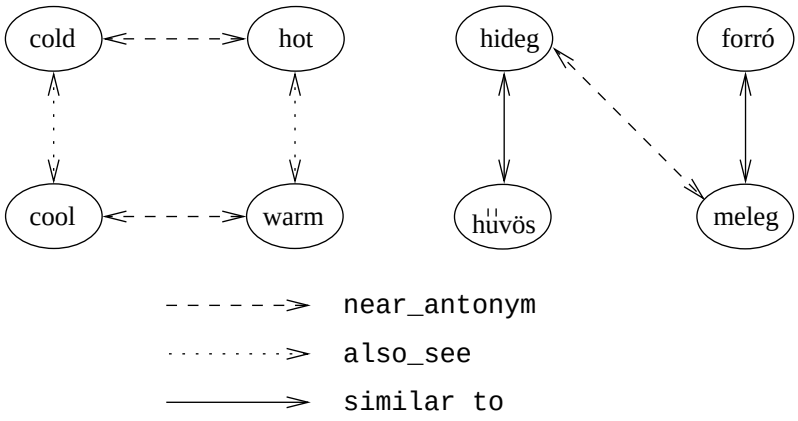

Fig. 13: Differences between the cold-warm domain

On the conceptual level, there is only one antonym pair in this dimension: *cold* and *warm*. However, on the lexical level, there are two oppositions in English: *cold-hot* and *cool-warm*. All of the four words can be matched with a Hungarian equivalent (*hideg, forró, hűvös, meleg*), but the relations cannot be automatically applied to Hungarian since in Hungarian, there is only one antonym pair. The supposed *antonym* relation *hideg-forró* could point only from *forró* to *hideg* since a Hungarian speaker would associate *forró* with *hideg* (or *jéghideg* 'stone-cold') while *hideg* with *meleg*.

There are differences in the lexicons of the two languages as well. Some adjectives of PWN cannot be paired with a Hungarian equivalent. This can be due to two reasons. First, the concept is not lexicalized in Hungarian, thus, there is no such adjective (*unattractive – nem vonzó* lit. not attractive, which is a non-lexicalized synset in HuWN). Second, the concept expressed by an English adjective can be lexicalized by a word belonging to a different part-of-speech in Hungarian: *afraid* (adjective) – *fél* (verb). In order to mark these matchings, the new relation *eq_xpos_synonym* has been introduced, which signals synonymy between different parts-of-speech.

When inserting a new word into the lexical database, some considerations should be taken into account. On the one hand, it is necessary to make sure that the word really exists, thus it is justified to include it in the dictionary. On the other hand, the part-of-speech of the word should be

determined. In these investigations traditional dictionaries such as The Concise Dictionary of Hungarian (ÉKSz) were of great help. However, the lack of occurrence does not necessarily mean the exclusion of the word – in this case, data from corpora (e.g. Hungarian National Corpus (Váradi 2002) can also influence the decision.

Determining the part-of-speech may be problematic even when the word occurs in a traditional dictionary. There are some tests to distinguish between adjectives and other parts-of-speech, which may prove to be useful for lexicologists (Kiefer 2006, Komlósy 1992).

For instance, tests differentiating between adjectives and participles include but are not limited to:

- only adjectives can be predicative: *ez a hír megdöbbentő* 'this piece of news is shocking' vs. **ez a hír Pétert megdöbbentő* lit. this piece of news is Peter shocking 'this piece of news shocked Peter'.
- Adjectives cannot preserve the arguments of the verb (excluding adjuncts): *a Pétert megdöbbentő hír* ' the piece of news that shocked Peter' vs. **ez a hír Pétert megdöbbentő* lit. this piece of news is Peter shocking 'this piece of news shocked Peter'.
- Only adjectives can be compared, participles cannot: **Pétert megdöbbentőbb hír* ' the piece of news that shocked Peter more intensively'

For more tests see Komlósy (1992).

As it can be seen, tests are not always reliable, they can only reveal certain tendencies. The word *alvó* can refer to a noun 'bedroom' and a participle 'someone who is sleeping' according to ÉKSz. As used of volcanoes, it is a lexicalized adjective, however, based on the tests, it could not be qualified as an adjective. Our decision is supported by the following test that helps to identify lexicalized attributive constructions (e.g. *vágott virág* 'cut flower'):

- the meaning of the construction is specific (*vágott virág* is not equivalent to a flower that is/was cut (into pieces)
- the modified word loses its main stress ('*vágott virág* and not '*vágott* '*virág*)
- Further attributes can modify only the whole structure, thus, they cannot intervene between the adjective and the noun nor can they modify only the noun *(*vágott rózsaszín virág* 'cut pink flower',**rózsaszín, vágott virág* 'pink, cut flower')
- The structure cannot be transformed into a predicative construction *(*A vázában levő virág vágott*. ' The flower in the vase is cut.')

On the basis of this, the inclusion of *alvó* as an adjective can be justified since the expression *alvó vulkán* 'dormant volcano' is a lexicalized attributive construction. Hungarians do not say either **A szigeten álló vulkán alvó* 'the volcano on the island is dormant' or **alvó nagy vulkán* 'dormant big volcano' and the construction forms one unit as far as stress is concerned.

Based on a similar argumentation, the adjective *ázott* 'wet'was also included into HuWn, however, it is not present in either ÉKSz or A magyar nyelv nagyszótára (Dictionary of the Hungarian Language).

5.2.3. Atypical dimensions

Some descriptive adjectives do not fit into the typical bipolar cluster structure of PWN. They occur in clusters having more focal synsets than the usual number, i.e. more than two adjectives are meant to express opposing values of an attribute, see Figure 14.

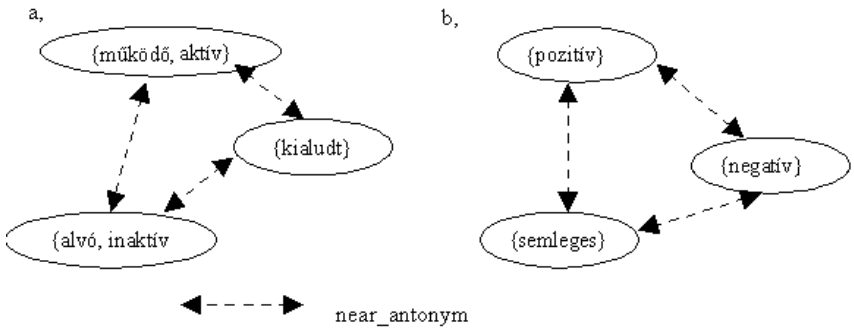

Fig. 14: Atypical adjectival clusters

The focal synsets of these domains form a „triangle" along the *near_antonym* relations running between each pair among them. Considering this representation, it might be deduced that these attributes are not bipolar but are of 3 dimensions, having three marked "poles". In the present section we argue for an alternative kind of representation, which, with the help of two new relations, enables adherence to the original bipolar structure of adjective clusters.

Descriptive adjectives are organised in clusters along semantic similarity and antonymy between words (instead of concepts), reflecting psychological principles. Consider the example in Figure 15. The adjective pair *pozitív* 'positive'-*negatív* 'negative' are the opposing poles of their domain. The situation of the word *semleges* 'neutral' is odd. Its English equivalent occurs as a third focal synset in the same domain as *positive* and *negative* in PWN. Relying on word association tests for Hungarian, we did not follow the solution of PWN when inserting *semleges* ('neutral') into HuWN. While the words *pozitív* and *negatív* do evoke each other in word association tests, the relation between *pozitív* and *semleges*, and *negatív* and *semleges*, respectively is not as straightforward. Although the word *semleges* does evoke *pozitív*, the antonym pair of *pozitív* is the adjective *negatív*. Loosening the scope of the usage of the relation *near_antonym* in order to enable antonym triangles to fit into a wordnet might cause anomalies in regular bipolar clusters as well (cf. direct and indirect antonyms). Therefore we have defined a new relation as an alternative to dealing with the case of triangles described above.

The adjectives *pozitív* and *negatív* determine a bipolar domain. This domain differs from the typical domains in the number and structure of its members. Apart from the two focal synsets, there is another adjective whose role is marked, but, as we have already shown, it is no real antonym of neither *pozitív* nor *negatív*. Furthermore, this special adjective expresses a value lying exactly in the *middle* of the domain. Therefore, the new relation we used in HuWN is called *middle*, and points to both focals of the given domain (Fig.15.).

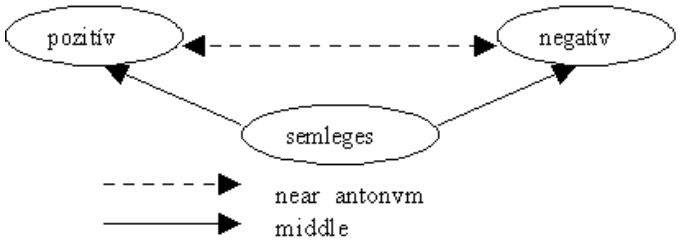

Fig. 15: The *middle* relation

It should be noted that the newly introduced relation *middle* can be used in any bipolar domain where the exact value (either being actually or considered conceptually as a discrete point) is lexically marked, e.g. in the domain determined by the adjectives *alsó-felső-középső* ('lower-upper-middle'). Although we have defined *middle* in relation to HuWN, it may be used in other wordnets, as well, since the above described case is not limited to the Hungarian language alone.

At first sight the scalar *middle* relation could be used in the example shown in Figure 14. The two opposing poles of the domain are {működő, aktív} 'active' and {kialudt} 'extinct, inactive', while the midpoint is denoted by {alvó, inaktív} 'dormant, inactive'. In this domain, however, the

middle value of the attribute cannot be considered as discrete. Furthermore, the synset {alvó, inaktív} might be considered to be in *similar_to* relation with {működő, aktív}, as the adjective *alvó* 'dormant' refers to a "presently not functioning volcano", thus having a closer meaning to {működő, aktív}, just as *langyos* 'lukewarm' is in *similar_to* relation with *meleg* 'warm'.

The domain specified by these three synsets differs from the aforementioned domains not only because of the similarities and contrasts between its members. These adjectives also constrain their scope: they can only refer to *volcanos*, and the wordnet has to account for this semantic relation. PWN and BalkaNet relate these adjectives through the use of the antonymy relation, and do not even indicate the relation with the noun exclusively modified by these adjectives from one point of view.

The synset-triple concerning volcanos is not the only triangle of this kind present in the semantic lexicon. For another simple example, we refer to the adjectives *egynyári-kétnyári-évelő* 'annual, biennial, perennial'. Had we only the *near_antonym* relation at our disposal, the information that the respective adjectives can only refer to plants would have to be omitted, and the fact that these three adjectives belong together could indeed only be present in a triangle form among them.

When taking a closer look, one can see that the adjectives mentioned above *partition* the extension of the particular noun, i.e. they divide the set of nouns, e.g. all the plants in the last example, into disjoint subsets. This motivates the name of the suggested new relation: *partitions*, which is represented as a pointer pointing from the adjectives to the noun synset they partition (see Fig. 16.).

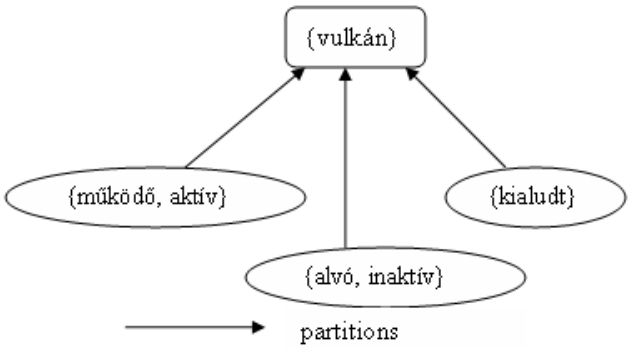

Fig. 16: The *partitions* relation

With the introduction of this new relation the explicit designation of the opposition between the adjective synsets becomes redundant, since due to the nature of the partitioning relation they may only be mutually exclusive. Although the *partitions* relation is similar to the *category_domain* relation of the wordnet, the two relations should not be confused. *Category_domain* relates the given adjectival meaning and the domain it can be used in, e.g.: {egyvegyértékű, monovalens} 'monovalent '– {kémia, vegyészet} 'chemistry', but does not specify the noun(s) it can modify, even if it can modify a certain noun exclusively.

6. 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.

7. Domain ontologies

7.1. 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. Nowadays, one of the most dynamically developing areas is the domain of finance and business, which makes heavy demands on applications in language technology. The importance of communication between business partners with different native languages can hardly be overestimated since Hungary became a member state of the European Union. The sudden increase in the quantity of business news requires the constant development of information extraction tools designed for this domain. Domain ontologies specifically tailored to the special terminology of a domain can serve as a basis for information extraction systems.

The financial domain ontology connected to the general HuWN ontology served as a basis for information extraction application.

To construct a business domain ontology, first of all the typical terms used in business communication must be identified. When collecting these terms, our group made use of two different strategies.

First, our linguists read business and financial news on the one hand and websites on political and economic issues on the other. They scanned these texts for business term candidates, which were collected into lists based on their part-of-speech. In order to avoid superfluous homogeneity, two different domains were selected for collecting terms. One is the Short Business News Corpus developed by us. Its data can be seen in Table 3.

Subject code	**Topic name**	**Number of pieces of news**
04006009	Stock market	2122
04008001	Central banks	300
04008004	Economic indicators	300
04008006	Exchange market	301
04008009	Interest rates	300
04009004	Agricultural and raw material market	300
04016003	Annual reports	310
04016005	Merge, acquisition, change of holder	316
04016012	Company descriptions	309
04016016	Income forecasts	301
04016018	Incomes	300
04016027	Court procedures and regulations	300
04016033	Opening of a new factory	300
04016034	Privatization	301
04016038	Quarterly or semi-annual economic reports	456
	Total	**6516**

Table 3: Topics of business news

As for the second source of data, the non-forum-like websites of the domain www.magyarorszag.hu were selected. Elements of the lists were transformed into synset candidates automatically, and the linguists in our group then decided whether or not to include them in the domain ontology. If the synset was already present in the general ontology, it was obviously disregarded; that is, it was not duplicated. If the synset candidate was to be included in the economic subontology, it was linked to its English equivalent in PWN 2.0 (if any), and it was inserted into the already existing hierarchy. The number of potential synsets is shown here:

Part-of-speech	**Number**
Noun	2835
Adjective	270
Adverb	6
Verb	181
Total	**3292**

Table 4: The number of potential synsets

7.1.1. Verbs in the financial domain ontology

The inclusion of verbs into the domain ontology was carried out within the frame of the general principles described earlier (see 4.). In order to develop the IE system, 69 verbs were introduced in the ontology, resulting in 86 synsets. Now, the domain ontology covers 222 + 216, that is, 440 literals (381 verbs in 114 hyponym synsets). From the corpus www.magyarorszag.hu, 181 verbal senses were added.

7.1.2. Borrowing financial terms from PWN

Besides collecting terms from corpora, we made use of PWN synsets when extending our financial domain ontology. By manual inspection, we located 32 concepts in PWN that we found to contain relevant terms in the domains of economy, enterprise and commerce. This strategy sought to provide more complex encyclopedic knowledge in this field. These concepts and their hyponyms (that is, their subtrees) were then automatically translated into Hungarian, transformed into synsets and then checked manually by our linguists. The ID numbers and synonyms belonging to these synsets and the number of (indirect) hyponyms are presented here:

ID	**PWN**	**HuWN**	**# hyponyms**
ENG20-06118498-n	contract:1	szerződés:1	32
ENG20-12486528-n	ownership:1	tulajdon:2, birtok:4, birtoklás:1, tulajdonjog:1, tulajdonlás:1	10
ENG20-01043364-n	transaction:1, dealing:2, dealings:3	lebonyolítás:1	200
ENG20-01056649-n	payment:2, defrayal:1, defrayment:1	kifizetés:1	14
ENG20-07857433-n	economy:1, economic system:1	gazdaság:3, gazdasági rendszer:1, gazdasági rend:1	18
ENG20-05780838-n	economics:1, economic science:1, political	közgazdaságtan:1	6

	economy:1		
ENG20-09401295-n	economist:1, economic expert:1	közgazdász:1	37
ENG20-12637385-n	liabilities:1	tartozás:2	79
ENG20-12571125-n	financial loss:1	anyagi kár:1, veszteség:1	252
ENG20-12520120-n	cost:1	költség:2, összköltség:1	233
ENG20-07565031-n	financial institution:1, financial organization:1, financial organisation:1	pénzintézet:1, pénzügyi szervezet:1, pénzügyi intézmény:1	57
ENG20-01044450-n	transfer:6, transference:2	átruházás:1	10
ENG20-07566541-n	enterprise:2	vállalkozás:4, vállalat:1	121
ENG20-01031794-n	commercial enterprise:2, business enterprise:1, business:2	gazdasági vállalkozás:1	108
ENG20-07571175-n	business:1, concern:3, business concern:1, business organization:1, business organisation:1	üzleti szervezet:1	75
ENG20-07567480-n	agency:2	ügynökség:2	11
ENG20-07570097-n	firm:1, house:6, business firm:1	cég:1	26
ENG20-07569639-n	corporation:1, corp:1	bejegyzett cég:1	13
ENG20-01035703-n	finance:1	pénzügy:2	12
ENG20-07575208-n	commercial enterprise:1	kereskedelmi vállalkozás:1	40
ENG20-07568361-n	company:1	cég:1, vállalat:1, társaság:6	50
ENG20-07572756-n	publisher:1, publishing house:1, publishing firm:1, publishing company:1	kiadóvállalat:1	3
ENG20-01028287-n	commerce:1, commercialism:1, mercantilism:2	gazdasági tevékenység:1	175
ENG20-09007401-n	consumer:1	fogyasztó:2	69
ENG20-09253155-n	businessperson:1, bourgeois:1	tőkés:2, burzsoá:1	189
ENG20-01046774-n	deal:1, trade:5, business deal:1	üzlet:3	3
ENG20-01049567-n	selling:1, merchandising:1, marketing:1	árusítás:1, árulás:1, eladás:1	22
ENG20-00073027-n	trading:1	kereskedelem:2, kereskedés:2	6
ENG20-01032803-n	business activity:1, commercial activity:1	üzleti tevékenység:1	7
ENG20-09861061-n	salesperson:1	eladó:4, elárusító:1	10
ENG20-03607786-n	mercantile establishment:1, retail	üzlet:2	70

	store:1, sales outlet:1, outlet:1		
ENG20-03583390-n	marketplace:2, mart:1	kereskedelmi központ:2	13
Total:			1971

Table 5: Financial terms in PWN and HuWN

1206 synsets are available form the 32 root concepts (those that can be reached in several different ways are counted only once). Out of these, 266 synsets were already translated in an earlier phase of the project, thus, it was only 940 synsets that were translated following the usual protocol. Machine translation heuristics provided possible Hungarian equivalents for 356 synsets.

7.1.3. The prototype of the business information extraction system

Semantic frames supporting ontology based information extraction

Semantic features occurring in the frames are changed to synsets covering the senses encoded in the features. Although the frames still contain the semantic features, a submodule of the program changes them to the appropriate synsets. Thus, should the sense number of a literal in the synset change in a later stage, the change should be carried out only once (in the program) and there is no need to change all the semantic frames including that literal. The correspondence between synsets and semantic features are shown below:

Animate:
ENG20-00003009-n (élőlény:1/living thing:1, animate thing:1)
ENG20-00004824-n (sejt:1/cell:2)

Human:
ENG20-00006026-n (ember:1, egyén:1, emberi lény:1, halandó:1, személy:1, valaki:1, lélek:1/ person:1, individual:1, someone:1, somebody:1, mortal:1)

Abstract:
ENG20-00020333-n (mentális jelenség: 1/psychological feature:1)
ENG20-00020486-n (elvont fogalom:1, absztrakció:1/abstraction:6)

Bodypart:
ENG20-04919813-n ((szervezet alkotórésze):/body part:1)

Measure:
ENG20-12810936-n (alapmértékegység:1/fundamental quantity:1, fundamental measure:1
ENG20-12833460-n (hosszmérték:1/linear measure:1, long measure:1)
ENG20-12811168-n (meghatározott mennyiség:1, adott mennyiség:1/definite quantity:1)
ENG20-12812220-n (mértékegységrendszer:1/system of weights and measures:1)

Dynamic:
BCSHu-2020168512 (szerkezet:8)
ENG20-02988377-n (szállítóeszköz:3/conveyance:3, transport:1)
ENG20-03633712-n (mixer:4/mixer:4)
ENG20-03706018-n (optikai eszköz:1/optical device:1)
ENG20-03158939-n (elektronikus eszköz:1/electronic device:1)
ENG20-04100622-n (fényforrás:1/source of illumination:1)
ENG20-02754218-n (fúvóeszköz:1/blower:1)
ENG20-03857090-n (projectile:1, missile:2)
ENG20-04107553-n (dárda:1, lándzsa:1/spear:1, lance:1, shaft:7)

ENG20-03222124-n (kézifegyver:1/firearm:1, piece:7, small-arm:1)
ENG20-03706957-n (optikai műszer:1/optical instrument:1)
ENG20-03185523-n (robbanóeszköz:1/explosive device:1)
ENG20-03846203-n (elektromos szerszám:1/power tool:1)
ENG20-03293100-n (kerti szerszám:1/garden tool:1, lawn tool:1)
ENG20-10687119-n (atmoszferikus jelenség:1, légköri jelenség:1/atmospheric phenomenon:1)
ENG20-03398495-n (háztartási gép:1/home appliance:1, household appliance:1)
BCSHu-1439559362 (gép:8)

Company:
ENG20-07523126-n (szervezet:3, organizáció:4/organization:1, organisation:3)

Time:
ENG20-00023548-n (idő:1/time:5)
ENG20-14367213-n (idő:3/clock time:1, time:6)
ENG20-14296945-n (időegység:1/time unit:1, unit of time:1)

Mass:
ENG20-00017572-n (anyag:1/substance:1, matter:1)
ENG20-13935705-n (hatóanyag:2/agent:2)
ENG20-08869095-n (építőelem:2, elem:2/unit:5, building block:1)

Currency:
ENG20-12615184-n (valuta:1/medium of exchange:1, monetary system:1)
ENG20-12627781-n (valuta:1 /monetary unit:1)

Weather:
ENG20-14375231-n (évszak:1/season:2, time of year:1)
ENG20-10782227-n (időjárási körülmény:1/weather:1)
ENG20-10707446-n (kondenzáció:1, páralecsapódás:1/condensation:3, condensate:1)

Now, 396 semantic frames have an equivalent in the new format.

7.2. The Hungarian legal wordnet

The first steps towards a general legal wordnet for Hungarian have been taken since we have constructed an ontology of concepts related to financially liable offences (customs law wordnet (TaXWN).

After having created the hierarchy of concepts, possible ways to join an international legal wordnet called LOIS were examined. First, synsets and concepts of TaXWN and the English version of LOIS were contrasted, then the IDs of corresponding synsets were inserted into the Note slot of Hungarian legal synsets. Thus, with the help of the interlingual index, Hungarian legal synsets are matched to those in LOIS. At the moment, this correspondence exists only from TaXWN synsets to LOIS synsets but not vice versa.

The quality and quantity of the ontology fulfilled the initial expectations and it can offer a theoretical and empirical base for a future legal wordnet covering other legal topics.

In the framework of the customs law WordNet project, the researchers from Szeged first began to collect a term vocabulary from Hungarian legal texts by automatic methods. The consortium finally decided that two acts should be processed: Act on taxation procedure[3] and Act on excise duty[4]. Legal experts from the Department of Constitutional Law were invited to the

[3] Hungarian Act no. XCII. of 2003.
[4] Hungarian Act no. CXXVII. of 2003.

project. They manually checked the terminology and advised to augment them with other important terms e.g. from the Penal Code. Unfortunately, they had no other digitized resource to begin with. Later the consortium asked the researchers from Szeged to add further terms from the publicly available commands of the Commissioner. When the list of terms was finalized, legal experts began to collect glosses. The related laws, decrees and legal handbooks were systematically thumbed over. If more than one gloss was found for a term, then all explanations – having made a record of their source – were included in the knowledge base.

When the term vocabulary was finished, computational linguists together with legal experts ordered the terms in a hierarchy. The originally paper-based notes and Microsoft Excel spreadsheets were compiled into a WordNet by linguists using the VisDic editor program (Horák and Smrž, 2004). Principally, the hypernymy relation was implemented but also holonymy occurred several times.

7.2.1. The LOIS Legal WordNet

The LOIS (Legal Ontologies for Knowledge Sharing) multilingual WordNet was created during an EU funded project EDC 22161 between 2003 and 2006 (Dini, Peters, et al. 2005, Peters, Sagri and Tiscornia 2007). The LOIS consortium was led by the Italian Institute of Legal Information Theory and Techniques in Florence. After a short negotiation a research agreement between the Institute of Informatics at Szeged and the LOIS consortium was signed according to which, Hungarian researchers were granted access to the LOIS multilingual legal WordNet.

The LOIS WordNet originally contained 35000 concepts in five European languages (English, German, Portuguese, Czech and Italian), roughly 7000 concepts in each.

```
<WORD_MEANING ID="1429"
 PART_OF_SPEECH="N" STATUS="FINISHED">
 <SOURCEBASE>LEXDB</SOURCEBASE>
 <NOTE/>
 <GLOSS>a person who has not reached full legal age</GLOSS>
 <CONCEPTS/>
 <VARIANTS>
   <LITERAL LEMMA="minor" SENSE="1">
    <EXAMPLES>not    of    legal    age;    &amp;quot;    minor
children&amp;quot;</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="minor" SENSE="1">
    <EXAMPLES>a person who has not reached full legal age; a
child or juvenile</EXAMPLES>
   </LITERAL>
   <LITERAL LEMMA="juvenile" SENSE="1">
    <EXAMPLES>a person who has not reached the age (usually
18) at which one should be treated as an adult by the criminal
justice system</EXAMPLES>
   </LITERAL>
 </VARIANTS>
</WORD_MEANING>
```

Fig. 17. The concept of *juvenile* as defined in the LOIS WordNet

The LOIS WordNet uses its own Inter-Lingual Indices to identify the concepts (synsets). The IDs of the semantically identical synsets are the same in each of the five languages. Synsets, mostly nouns, are taken from the general legal science and there are few verbs, adjectives and

adverbs. Generally, each synset has a definition which sometimes comes from Celex[5], the legal document repository of the EU or from legal handbooks. In Figure 17 an example of a LOIS synset is shown.

7.2.2. A synset in the legal wordnet

The <DEF> node (gloss) contains the definition of the synset, which legal experts usually took from an act being in force or from legal handbooks. The part-of-speech of the synset is marked in the <POS> node. Synonyms of a term were collected from legal handbooks. In several cases, synonyms were multiword expressions due to the characteristics of the legal terminology. Linguistic relations like hypernymy or holonymy were coded in <ILR> nodes. The <ID> nodes contain the ILI indices of the synsets.

In Figure 18 an example of a synset from the Hungarian customs law WordNet is shown. It can be seen, that the Hungarian counterpart of the LOIS synset "juvenile" has a Hungarian WordNet <ID> due to the fact that the customs law WordNet was made as an extension to the Hungarian WordNet.

In the first <SNOTE>, one can find the exact reference to the legal place where the gloss is taken from, namely Penal Code (Law IV. of 1978.), section 107. In the second <SNOTE>, the LOIS ILI index and an explanation in Hungarian are included.

```
<SYNSET>
  <ID>HuWN-911671085</ID>
  <SYNONYM>
  <LITERAL>fiatalkorú
     <SENSE>0</SENSE>
  </LITERAL>
  </SYNONYM>
    <DEF>Fiatalkorú az, aki a bűncselekmény elkövetésekor
tizennegyedik élet évét betöltötte, de a tizennyolcadikat még
nem.</DEF>
  <SNOTE>1978. évi IV. tv. Btk. 107.§. (1)</SNOTE>
  <SNOTE>LOIS ID="1429"; a magyar jogrendben kis- és
fiatalkorú megkülönböztetés létezik</SNOTE>
  <SNOTE>jog</SNOTE>
  <POS>n</POS>
  <ILR>HuWN-148541600
     <TYPE>hypernym</TYPE>
  </ILR>
</SYNSET>
```

Fig. 18: The concept of *fiatalkorú* (*juvenile*) as defined in the customs law WordNet

7.2.3. Conflicts between linguistic and legal requirements

When building the WordNet it was often found that the requirements of linguistics and law were contradictory so researchers had to make priorities. It was decided that, first, they meet the requirements of law and, then, take linguistics into consideration where possible.

[5] http://eur-lex.europa.eu/en/index.htm

As a consequence, the customary linguistic rule applied in WordNets that the definition of a synset must contain a hypernym of the concept or its synonym (Miller et al., 1990) has been modified for, in most cases, definitions are mere lists of words.

In the Hungarian WordNet (Alexin et al., 2006; Miháltz et al., 2008), within synsets, notes are units that make short, supplementary comments possible. However, in the customs law WordNet notes have been given a new function. They are used to include information that cannot be entered as a part of the definition but provide substantial, indispensable data e.g. exact place of the definition in the legal texts, numerical data (e.g. alcohol concentration, quantity of importable goods, etc.)

When creating the hierarchy, the *bottom-up* method was followed because concepts derived from legal sources proved to be rather specific and they were usually used to create base-level synsets only. This, however, made the work simpler because hypernyms could be selected relying on the hierarchy of Hungarian WordNet.

In the customs law WordNet there are nine *unique beginner* synsets. Due to the decision mentioned above, it may happen that an element identified as an object on the base-level gets linked to a non-object hypernym synset or occurs in the tree of the *unique beginners* e.g. *abstraction* or *state*. This linguistically indefensible state was impossible to eliminate. Due to the phraseology of law these apparent "inconsistencies" have remained.

7.2.4. Connections between the Hungarian customs law WordNet and the LOIS Legal WordNet

The last step of the work was to establish connections between the two WordNets. Legal experts examined the English version of the LOIS WordNet and produced a list of synsets that may have connections to the customs law WordNet. A linguist and a legal expert then – taking the definitions into consideration – checked manually the list item by item to figure out whether the relation between the two concepts is valid, It was also checked whether the LOIS synset was more general than the synset in the customs law WordNet. In several cases the LOIS WordNet did not contain glosses for the synsets therefore the decision on identicality could not be made.

When the two synsets proved to be undoubtedly identical, the connection has been marked in the *note* field of the synset in the customs law WordNet as follows: LOIS ID="nnnn", where nnnn is the ILI index of the corresponding synset in the LOIS WordNet. A short explanation was also added. See Figure 2.

	Connected to LOIS	Cannot be connected to LOIS	All
General legal synset	81	116	197
Excise duty synset	113	337	450
Total	194	453	647

Table 6: The number of connections between the customs law WordNet and the LOIS WordNet

In Table 6, statistics on the customs law WordNet is presented. 194 out of the 647 (30%) synsets from the customs law WordNet have a counterpart in the LOIS WordNet. Among them 113 synsets are closely connected to the excise duty terminology (declaration, payment, definitions, crimes etc.), while 81 synsets are general legal terms.

In the whole customs law WordNet, 450 out of the 647 synsets were taken from the excise duty terminology. Their definitions come from legal rulings (laws, decrees, orders, etc.) being in force, e.g. tax warehouse, licensee of the tax warehouse, the onset of tax paying obligation. The

remaining 197 synsets are general legal terms with definitions taken from handbooks, e.g. interest, loss, official, representation.

The number of adjectives, nouns and verbs in the two WordNets are shown in Table 7.

	LOIS WordNet (English)	Customs Law WordNet
adjectives	0	0
nouns	6720	647
verbs	51	0

Table 7: The distribution of the adjectives, nouns, and verbs among the synsets of the two WordNets

8. Conclusions

The project described here aimed to build the Hungarian wordnet based on international norms. Standards of wordnet constructions were followed in the construction process as far as possible, however, due to some language specific or language independent reasons, some novelties were also introduced and some new relations were invented. This was because some problems that occurred when building synsets could not be solved with relations traditionally used in wordnets. For this reason, some new relations were introduced:

- middle
- partitions
- is_consequent_state_of
- is_preparatory_phase_of
- is_telos_of

See 4.2.2. and 5.2. for more details.

For language-specific reasons, it proved to be necessary to represent the complex structure of events in HuWN. Thus, some new relations were invented and nucleus nodes were inserted into the hierarchy.

Since it is impossible to find a perfect overlap between the concepts of two languages, it is inevitable to have some synsets that can only be circumscribed in the other language. During the construction of HuWN, PWN functioned as the starting point, thus, in several cases, we were confronted with such synsets. These were marked as non-lex synsets.

A basic building principle of HuWn was that the parent synset and its child should not share any of their literals. In cases where the hyponym synset was lexicalized but in the same way as its hypernym (i.e. the same word form expressed both concepts), the hyponym was marked as t non-lex.

Problems

Problems concerning part-of-speech (numerals, pronouns)

There are some differences between Hungarian and English grammatical traditions concerning the part-of-speech of certain word classes. For instance, the English equivalents of those words that are classified as numerals in Hungarian dictionaries are adjectives in English dictionaries. This classification is reflected in PWN as well, however, according to ÉKSz., these words are numerals. Since in HuWN there are only four parts-of-speech (noun, verb, adjective, adverb), it was necessary to mark these synsets as *t non-lex* and the Note slot contains *más szófaj* ‘other part-of-speech’.

This problem involved not only numerals but certain pronouns as well (e.g. {other} vs. {más}).

Problems concerning part-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-spech used in wordnets. For instance, the English word *afraid* is an adjective, however, its Hungarian counterpart *fél* is a verb. In these cases, we made use of the relation *eq_xpos_synonym*.

Underdetermined relations

Some of the relations initially applied in PWN proved to be overgeneralizing or underdetermined. In order to correct these inconsistencies, several initiations were introduced or proposed: in EuroWordNet, holonymy-meronymy relations are extended and used in a somewhat different manner than in PWN (Alonge et al. 1998). About the division of the relation antonym see Vincze, Almási, Szauter 2008.

Interoperability of wordnets

Wordnets were originally planned to be a computational model the human lexical memory and launched by psycholinguists of the cognitive department of Princeton University. From a computational point of view, wordnets are massive and well-structured databases in which thousands of words and meanings are organized into a semantic network.

Wordnet projects (PWN, EuroWN, BalkaNet, HuWN) aimed at providing a connection between databases of different languages with the help of the so-called interlingual index.

Interoperability makes it possible to gain the same information from different languages at the same time (multilingual information extraction). It may also prove useful in editing bilingual dictionaries, extracting information from multilingual resources, and enhancing the quality of machine translations. In order to examine the potential applications, the following calculations were performed.

We compared the wordnet databases of ten languages – Bulgarian, Czech, German, Estonian, English, Spanish, French, Hungarian, Italian and Dutch – and examined how much they overlap each other.

First of all, we investigated the number of common synsets of two languages, that is, how many concepts they share. Results show that the overlap between small wordnets is relatively big. In the case of larger databases, the overlap seems to be incidental: it cannot be stated that those wordnets share most of their concepts with PWN (which functioned as their model for construction) – see the case of Dutch and Italian wordnet. More striking is the identity of English and Bulgarian and English and Czech wordnets. It virtually means that all of the Bulgarian synsets were translated from English, that is, each Bulgarian synset has an English counterpart. It also entails that there are no synsets that are directly from Bulgarian – that is, without an English equivalent. The situation is somewhat similar in the case of Czech, however, 48 synsets can be found that are originated from Czech (there is no English counterpart for them). These data might also reveal the strategies used when constructing these wordnets.

Based on the results, the overlap between synsets of different wordnets can be considered relatively insignificant, however, in the case of Bulgarian-English (100%), Czech-English, Spanish-English and French-English language pairs it is very high. The number of synsets occurring in all the languages examined is only 292 although the indices in languages are above 10000. These data strongly undermine our expectations concerning multilingual applications.

To sum up, the following conclusions can be drawn:

The original aim to ensure interoperability between ontologies of languages in the project could be fulfilled only minimally.

The joint use of wordnets is seriously undermined in this way.

It is also questionable whether a wordnet heavily (or completely) relying on PWN can be considered as a conceptual network representing the given language (see e.g. the Bulgarian WordNet).

There can be significant differences in the conceptual networks of two languages, and it is dubious whether they can be represented by the same conceptual structure.

As for a future extension of wordnets it is worth considering the inclusion of synsets that occur in most languages, in this way, the interoperability of wordnets might be improved.

The frequency of non-lex synsets also reflects the difficulties of matching concepts belonging to different languages. On the other hand, some of the concepts existing only in one wordnet are language- or culture-specific, that is, they cannot (or hardly can) be expressed in another language. (Obviously, the other part of such synsets could be expressed in other languages, however, at the moment, they are simply not included in the database.) Thus, it is preferable not to translate one wordnet as a whole into another language since the result will reflect the conceptual network of the source language and not the one of the target language.

Inconsistent use of *usage_domain*

In HuWN, the relation *usage_domain* was not applied because of the inconsistencies found in PWN: sometimes the usage label of a synset does not hold for all literals and this way of representation does not reveal which literals it is valid for.

Further plans and possibilities

The Hungarian wordnet makes it possible to develop other domain ontologies and to construct further NLP applications. In the medium term, the development realized in this project may enable the consortium members to participate in further European R+D projects – possibly with other (international) partners. The project may contribute to the development of multilingual applications based on a homogeneous European system.

References

Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy. 2006. Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction, Project report, In: *Proceedings of the Third International WordNet Conference GWC 2006*, South Jeju Island, Korea, 2006, pp. 291–292.

Alonge, A., Bloksma, L., Calzolari, N., Castellon, I., Marti, T., Peters, W., Vossen P.: The Linguistic Design of the EuroWordNet Database. J. Computers and the Humanities. Special Issue on EuroWordNet, 32(2–3), 91–115 (1998)

Bach, E.: The Algebra of Events. Linguistics and Philosophy 9 (1986) 5-16.

Dini, L., Peters, W., Liebwald, D., Schweighofer, E., Mommers, L., and Voermans, W. 2005. Cross-lingual legal information retrieval using a WordNet architecture. In *Proceedings of the 10th international Conference on Artificial intelligence and Law* (Bologna, Italy, June 06–11, 2005). ICAIL '05. ACM, New York, NY, 163–167.

Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)

Horák, A., Smrz, P. 2004. VisDic — Wordnet Browsing and Editing Tool, In: *Proceedings of the Second International WordNet Conference GWC 2004*, pp. 136-141.

Kiefer, F.: Aspektus és akcióminőség. Különös tekintettel a magyar nyelvre. [Aspect and Aktionsart. with Special Respect to Hungarian]. Akadémiai Kiadó, Budapest (2006)

Komlósy András 1992. Régensek és vonzatok. In: *Kiefer, F. (ed.): Strukturális magyar nyelvtan I. Mondattan*. Akadémiai Kiadó, Budapest, Hungary.

Miháltz, Márton: Constructing a Hungarian Ontology Using Automatically Acquired Semantic Information. *Proceedings of the IWCS-5*, Tilburg, The Netherlands (2003)

Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T. 2008. Methods and Results of the Hungarian WordNet Project, In: *Proceedings of the Fourth Global WordNet Conference. GWC 2008*, University of Szeged, Department of Informatics, 2008, pp. 311–320.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an On-line Lexical Database. J. International Journal of Lexicography 3(4), 235–244 (1990)

Moens, M., Steedman, M. Temporal ontology and temporal reference. Computational Linguistics, 14(2) (1998) 15–28.

Peters, W., Sagri, M. and Tiscornia D. 2007. The structuring of legal knowledge in LOIS, *Artificial Intelligence and Law*, Volume 15, Issue 2 (June 2007), pp. 117–135. Springer Verlag, ISSN: 0924-8463.

Prószéky, G., Miháltz, M., Nagy, D. 2001. Toward a Hungarian WordNet, *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June 2001.

Tufiş, D. (ed.), *Romanian Journal of Information Science and Technology. Special Issue on BalkaNet*, Vol. 7, No. 1–2, 2004.

Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas, pp. 385–389 (2002)

Veronika Vincze, Attila Almási, Dóra Szauter 2008. Comparing WordNet Relations to Lexical Functions. In: Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, Piek Vossen (eds.): *Proceedings of the Fourth Global WordNet Conference. GWC 2008*. Szeged: University of Szeged, Department of Informatics, 462–473.

Appendix

A summary of relations applied in HuWN

also_see	Adjectival focal synonym synset
be_in_state	Noun belonging to the adjective
category_domain	category
causes	causing
derived	Derived form
eng_derivative	Derived form (in English)
holo_member	member
holo_part	part
holo_portion	material
hypernym	hypernymy
is_consequent_state_of	consequency
is_preparatory_state_of	Previous state
is_telos_of	Culmination point
middle	middle
near_antonym	antonymy
near_synonym	synonymy
partitions	What noun it can modify
region_domain	location
similar_to	Synonym satellite adjective
subevent	subevent
temporal_precondition	Precedes in time
usage_domain	Usage domain
verb_group	Verb group (based on English)

Statistical data

	noun	**adjective**	**adverb**	**verb**	**total**
Number of synsets	**33530**	**4112**	**1039**	**3607**	***42288***
Non-lex synsets	**943**	**699**	**137**	**219**	***1998***
T non-lex synsets	**150**	**42**	**0**	**261**	***453***
Total numbers of literals	**45508**	**6215**	**1793**	**6947**	***60463***
Literal/synset rate	**1.36**	**1.51**	**1.73**	**1.93**	***1.43***

The following table represents the percentage rate of the above features.

	noun	**adjective**	**adverb**	**verb**
number of synsets	**79,28963299**	**9,723799**	**2,456962**	**8,529607**
Non-lex synsets	**47,1971972**	**34,98498**	**6,856857**	**10,96096**
T non-lex synsets	**33,11258278**	**9,271523**	**0**	**57,61589**
Total numbers of literals	**75,26586507**	**10,27901**	**2,96545**	**11,48967**

Synsets with ENG20 ID: 26369 62.36%
Synsets with HuWN ID: 15919 37.64%

The highest number of literals within one synset

n: 18 literals
a: 17 literals
b: 8 literals
v: 17 literals

The number of synsets with hypernym(s): 36586

Having 1 hypernym: 35532 synsets
Having 2 hypernyms: 976 synsets
Having 3 hypernyms: 67 synsets
Having 4 hypernyms: 10 synsets
Having 5 hypernyms: 1 synset

Synsets having at least 3 hypernyms are named entities (with the exception of {karácsony:1} 'Christmas').

ilr types

hypernym: 37730
similar_to: 6966
holo_part: 3344

near_antonym: 1745
holo_member: 1268
category_domain: 929
verb_group: 816
be_in_state: 412
also_see: 399
holo_portion: 132
region_domain: 131
is_telos_of: 112
is_preparatory_phase_of: 104
usage_domain: 100
causes: 92
subevent: 92
near_synonym: 77
is_consequent_state_of: 28
subevent_nec_of: 23
middle: 16
subevent_of: 16
temporal_precondition: 9
has_consequence: 8
converse: 4
aktionsart: 3
partitions: 2
near: 1