Sponsors

Research Academic Computer Technology Institute
University of Patras

Co-operating Organisations

Global Wordnet Association

The Workshop Programme

8:30 - 13:30 Morning Session: Wordnet Applications, Evaluation and Standardization

8:30 - 8:45 Welcome & Introduction; Overview about the workshop

8:45 - 9:30 **KEYNOTE SPEECH**
Christiane Fellbaum (University of Princteon):
Going global: Issues in the standardization of wordnets

9:30 - 10:00 Neeme Kahusk (University of Tartu):
A Lexicographer's Tool for Word Sense Tagging according to Wordnet

10:00 - 10:30 Piklu Gupta (Fraunhofer Institut Darmstadt):
Approaches to Checking Subsumption in GermaNet

10:30 - 11:00 Graham Katz, Jahn-Takeshi Saito, Joachim Wagner, Philip Reuter, Sabine Reinhard & Michael Burke (University of Osnabrueck):
Evaluation of GermaNet: Problems using GermaNet for Automatic Word Sense Disambiguation

11:00 - 11:30 Coffee break

11:30 - 12:00 Karel Pala & Pavel Smrz (University of Brno):
Glosses in WordNet 1.5 and their Standardization/ Consistency
(The Exercise for BalkaNet)

12:00 - 12:30 Lothar Lemnitzer & Claudia Kunze (University of Tuebingen):
Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet

12:30 - 13:00 Tomas Pavelek & Karel Pala (University of Brno):
Wordnet Standardization from a practical point of view

13:00 - 13:30 Overall Discussion and Conclusion for 1st Part of the Workshop:
* standardization and compatibility guidelines
* application and evaluation scenarios envisaged
* future actions

13:30 - 14:30 Lunch break

14:30 - 19:30 Afternoon Session: Wordnet Structures and Applications for the less-studied Languages

14:30 - 15:00 **Introduction Speech**
Prof. Dimitris N. Christodoulakis (University of Patras):
Structures of Semantic Networks

15:00 - 15:30 Dan Tufis (Romania Academy) & Dan Cristea (A.I. Cuza University):
Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet

15:30 - 16:00 Gabor Proszeky (MorfoLogic) & Marton Mihaltz (Eotvos Lorand University, Budapest):
Automatism and User Interaction: Building a Hungarian Wordnet

16:00 - 16:00 Dimitris Avramidis, Maria Kyriakopoulou, George Kourousias, Sofia Stamou & Manolis Tzagarakis (University of Patras, *RA* CTI Patras):
Viewing Semantic Networks as Hypermedia

16:30 - 17:00 Coffee break

17:00 - 17:30 Ioannis - Dimitris Koutsoumpos, Manolis Tzagkarakis & Dimitris Christodoulakis (University of Patras, *RA* CTI Patras):
Requirements for Domain-Specific Wordnets

17:30 - 18:00 Kadri Vider (University of Tartu):
Notes about labeling semantic relations in Estonian Wordnet

18:00 - 18:30 Irina V. Azarova, Olga A. Mitrofanova, Anna Sinopalnikova & Ilya Oparin (State University St-Petersburg):
Building the Lexical Database for the Russian Language

18:30 - 19:00 Natalia V. Loukachevitch & Boris V. Dordov (Moscow State University):
Development and Use of Thesaurus of Russian Language RuThes

19:00 - 19:30 Overall Discussion and Conclusion for the 2nd Part of the Workshop

Workshop Organisers

Dimitris N. Christodoulakis, Patras University (Greece)
Claudia Kunze, Lothar Lemnitzer, University of Tuebingen (Germany)
Karel Pala, Masaryk University Brno (Czech Republic)

Workshop Programme Committee

Christiane Fellbaum, Princeton University (USA)
Piek Vossen, Irion Technology Delft (The Netherlands)
Kemal Oflazer, Sabanci University Istanbul (Turkey)
Jeroen Hoppenbrouwers, Tilburg University (The Netherlands)
Randee Tengi, Princeton University (USA)
Wim Peters, Sheffield University (GB)
Kadri Vider, University of Tartu (Estonia)
Julio Gonzalo, UNED Madrid (Spain)
Palmira Marrafa, University of Lisboa (Portugal)
Paul Buitelaar, DFKI Saarbruecken (Germany)
Andreas Wagner, University of Tuebingen (Germany)
Erhard Hinrichs, University of Tuebingen (Germany)
Simonetta Montemagni, University of Pisa (Italy)
Robert Ermers, Van Dale Data BV (The Netherlands)

Table of Contents

Workshop on Wordnet Structures and Standardization, and how these affect Wordnet Applications and Evaluation

Author Index

A Lexicographer's Tool for Word Sense Tagging According to WordNet

Neeme Kahusk

University of Tartu
Department of General Linguistics
Tiigi 78-204, 50410 Tartu, Estonia
nkahusk@psych.ut.ee

Abstract

This paper describes a Web-based tool for tagging word senses according to Estonian WordNet. The tool makes use of EuroWordNet import-export format that is converted into XML. The user interface is divided into three main parts that provide information about the word to be tagged: (1) context (2) morphological analysis and (3) entries in lexicon (WordNet). The tool is aimed to facilitate lexicographers' work with languages, where morphological information is important at word sense disambiguation. The advantages of the tool and problems met are discussed in detail.

1. Introduction

The task of tagging word senses is demanding for the lexicographers. They have to find the words to tag from the text, find is the word presented in the lexicon, is there an appropriate meaning for the word in the lexicon, and finally, assign the meaning to the word in question.

Estonian is an agglutinative language, rich of word forms. A lot of word forms are ambiguous, and before getting lemma some morphological analysis is needed.

In the very beginning, the linguists who did the job, had to edit a plain text file and write the appropriate sense of word after its morphological reading.

To carry out word sense disambiguation, lexicographer has to know what are the different senses of the word. Thatswhy (s)he needs to see at least, definition (gloss), and example(s) of usage, and one hyperonym. Up to now, the people who did the job, edited files with a simple text editor and used Polaris tool, that seriously limited the number of workplaces where the job could be done. This drawback, and the fact that editing a file, where one can see only one word on a line, followed by morphological analysis, is a potential source of errors, rose the need for a tool that would be more task-oriented and usable on client-server basis.

As a result of integration output of morphological analyser and an automatic WSD system for finding words not in the thesaurus and pre-selecting senses, a tool was created that makes word sense tagging more accurate and less time-consuming.

The lexicographer's tool is working in two stages, off-line (preparatory) and on-line.

By implementing the tool we have found several problems that were not noticed at manual file-editing process. The dividing into parts of speech is a bit different in wordnets and Estonian morphological tradition; morphology, syntax and semantics are more tightly connected to each other than one can suppose. An additional feature of tagging multi-word units is needed.

2. Preparatory stage

For off-line stage, the following data files are needed: (1) current thesaurus (in import-export format); (2) file to disambiguate by word senses — it should be analysed by Estmorf, and piped through fs2kym. To ensure that the analysed text file has correct format (each word must have exactly one analysis), a small vaidating script is applied to it.

During off-line or preparatory stage, current version of Estonian WordNet (EstWN) is converted into XML. Then semyhe is applied to the data in two runs: on first run, nouns are disambiguated, on second one, verbs are disambiguated.

2.1. Morphological analysis

Morphological analysis is carried out with Estmorf provided by Kaalep (1997). In its original form, Estmorf outputs for every word its structure (stem, affixes and suffixes), part of speech and inflectional categories.

```
pea
    pea+0 //_D_ //
    pea+0 //_S_ sg g, sg n, //
    pida+0 //_V_ o, //
    pida+0 //_V_ o, //
```

Figure 1: Output of Estmorf from word form 'pea'.

Declinable words are differentiated into following parts of speech: common nouns or substantives (_S_), proper nouns (_H_), adjectives with positive degree, comparative degree and superlative degree (_A_, _C_, and _U_ respectively), numerals (_N_ cardinal, _O_ ordinal), pronouns and acronyms (_Y_). Possible sets of inflectional categories are given on the same line, if they occur inside one paradigm (structure and part of speech). Figure 1 illustrates analysis of 'pea':

1. adverb ('soon'; uninflected),
2. noun ('head'; singular, genitive or nominative),
3. and 4. verb (two homonyms[1]: 'keep' and 'must', both imperative, the last one modal, but this analysis does not show such features).

[1] There are more meanings: in EstWN there are 13 senses of verb 'pidama', but they can divided into 2 groups — the modal (3 senses) and main (10 senses) ones.

```
pea
    pea+0 //_D_ //
    pea+0 //_S_ com sg gen //
    pea+0 //_S_ com sg nom //
    pida+0 //_V_ main imper pres ps2 sg ps af //
    pida+0 //_V_ main imper pres ps2 sg ps neg //
    pida+0 //_V_ main indic pres ps neg //
    pida+0 //_V_ mod imper pres ps2 sg ps af //
    pida+0 //_V_ mod imper pres ps2 sg ps neg //
    pida+0 //_V_ mod indic pres ps neg //
```

Figure 2: Output of Estmorf from word form 'pea' piped through fs2kym.

It turned out that the output of Estmorf is not very good for disambiguation purposes. At first, the Estmorf analysis line itself contains ambiguous readings (different inflectional categories, although being inside one paradigm). Second, in some cases, the differentiation into parts of speech is too detailed. The authors of Estmorf have developed a conversion program fs2kym that modifies the output. Unfortunately the last version of fs2kym is not fully documented yet, the output is pretty much the same as used by Puolakainen (2001) and Roosmaa et al. (2001) for morphological disambiguation based on constraint grammar and syntactic analysis.

In fs2kym output, substantives and proper nouns are tagged as '`_S_ com`' and '`_S_ prop`' respectively. So are numerals and ordinals, '`_N_ card`' stands for numeral and '`_N_ ord`' for ordinal. In the same way all adjectives are tagged as `_A_`, their degree is added with next token: '`_A_ pos`' for positive adjective, '`_A_ comp`' for comparative and '`_A_ super`' for superlative one. For verbs, fs2kym adds inflectional readings with all possible solutions. Figure 2 illustrates previous example analysed with Estmorf and piped through fs2kym:

1. adverb ('soon'; uninflected),
2. noun ('head'; singular, genitive),
3. noun ('head'; singular, nominative),
4. verb ('to keep; to consider'; main, imperative, present, 2. person, singular, personal, affirmative)
5. verb ('to keep; to consider'; main, imperative, present, 2. person, singular, personal, negative)
6. verb ('to keep; to consider'; main, indicative, present, personal, negative)
7. verb ('must; should'; modal, imperative, present, 2. person, singular, personal, affirmative)
8. verb ('must; should'; modal, imperative, present, 2. person, singular, personal, negative)
9. verb ('must; should'; modal, indicative, present, personal, negative)

2.2. Preliminary word sense tagging

Preliminary word sense tagging is done with Semyhe system, as described by Vider and Kaljurand (2001). The main idea of Semyhe is based on a similar system by Agirre and Rigau (1996), using distances between the nodes corresponding to the word senses in the WordNet tree and the density of the tree. Contrast to the Agirre and Rigau system, Semyhe disambiguates both, nouns and verbs. Nouns and verbs are disambiguated in two separate runs, as they do not share the same hyperonym-hyponym hierarchies.

Fs2kym-piped output of Estmorf serves as input for Semyhe. As our aim at present stage is generating a word-sense disambiguated corpus, the Estmorf output is disambiguated by hand, so every word has only one reading. Semyhe adds its output to Estmorf analysis, an example is given in Figure 3 (upper part). Semyhe analysis is added to substantives and main or modal verbs. After last piece of morphological info '@' is added, then lemma in dictionary form (singular nominative for substantives, supine affirmative illative for verbs). The last two fields are separated with colon, last number denotes number of senses found from EstWN, the last but one is sense number found by Semyhe. If Semyhe finds more than one possible analysis, then the alternatives are separated by number sign (#).

2.3. XML format

The import-export (i/e) format of a language wordnet in EuroWordNet is derived from GEDCOM standard (Louw, 1998). The GEDCOM format[2] itself is hierarchical, so the initial conversion into XML is rather simple.

The main idea in converting EWN i/e format into XML was simplicity of conversion and not well-formedness or size of resulting file. So the current version of XML format is a simple translation of GEDCOM-format into XML: node labels are translated into elements (with some exceptions explained below), and node contents are translated into values of attribute '`VALUE`'. If element name consists of multiple words, element will be built from first letters of label name (`PART_OF_SPEECH` will be `POS`). Still there are some labels in EWN format that would be ambigous at such conversion. They differ only by plural ending. Such labels are converted so, that the ending 'S' is added to every subword: e.g. `USAGE_LABELS` will be `<USLS>` and `USAGE_LABEL` will be `<UL>` (Figure 4).

[2] http://www.gendex.com/gedcom55/55gctoc.htm

```
Pidas
    pida+s //_V_ main indic impf ps3 sg ps af // @ pidama:6:12
veidi
    veidi+0 //_D_ //
aru
    aru+0 //_S_ com sg part // 1 @ aru:1:1
ja
    ja+0 //_J_ crd //
lisas
    lisa+s //_V_ main indic impf ps3 sg ps af // @ lisama:3:3
liipsukese
    liipsu=ke+0 //_A_ pos sg gen //
liha
    liha+0 //_S_ com sg gen // @ liha:1:3
.
    . //_Z_ Fst //
-------------------------------------------------------
<s>
    <head id="1630" lemma="pidama" pos="V" class="main"
    rest="indic impf ps3 sg ps af" noofsenses="12" semyhe="6">Pidas</head>
    <other id="1631" pos="D">veidi</other>
    <head id="1632" lemma="aru" pos="S" class="com" rest="sg part"
    noofsenses="1" semyhe="1">aru</head>
    <other id="1633" pos="J" class="crd">ja</other>
    <head id="1634" lemma="lisama" pos="V" class="main"
    rest="indic impf ps3 sg ps af" noofsenses="3" semyhe="3">lisas</head>
    <other id="1635" pos="A" class="pos">liipsukese</head>
    <head id="1636" lemma="liha" pos="S" class="com"
    rest="sg gen" noofsenses="3" semyhe="1">liha</head>
    <other id="1637" pos="Z" class="Fst">.</other>
</s>
```

Figure 3: Upper: Output of Semyhe. Lower: The same sentence in XML format. Explanations in text. The analysed sentence can be translated like '[she] considered a bit and added a little slice of meat.'

The format will do for simple tasks like converting the thesaurus to form needed for literal browsing, but is not very suitable for more general tasks and is definitely not a good human-readable one.

There are several versions of EWN in XML that are more readable: (Kunze and Lemnizer, 2002; Smrz, 2002; Dowdall et al., 2002), and there is a special tool for viewing and editing WordNet in XML format: VisDic (Pavelek and Pala, 2002).

2.4. Text File in XML

The text file is also converted into XML. The format is similar to the one that was used at Senseval-2 task and training files, with some modifications. The `<sat>` elements are omitted (see sec. 4.3.), and `<other>` element is introduced for words being not heads, and for other tokens (punctuation marks). Morphological information is given as attributes for `<head>`[3]: `lemma, pos, class` and `rest`, the last one for other morphological reading. The identification number (position of token in text) is given as `id` attribute. Semyhe adds more attributes: `noofsenses` for number of senses in EstWN, `semyhe` for Semyhe applied sense number (Figure 3, lower part). Finally, the sense number assigned by lexicographer, will be inserted as value of `sense` attribute (not shown in the figure).

3. User Interface

After entering his/her name and selecting file to work with, user can move to main interface of the program.

The program window is divided into four frames: the main frame for text being analysed, morf frame and thesaurus frame. The lowest frame is for entering comments. In the uppermost frame user can browse text, words to disambiguate are in boldface, and depending on browser settings, underlined. Each word to disambiguate is preceeded by an identification number for references in comments.

User has to select appropriate sense for each word that needs disambiguation (these words are emphasized in bold and linkable). By clicking on appropriate word, user can see morphological information about the word (part of speech and word class), and thesaurus entries. The thesaurus entries are presented in a table: each row represents one synset. The 2nd column of the table shows members of synsets with sense numbers. In the 3rd column, there are explanations (glosses), and in last column there are ex-

[3] `id, pos, class` and `rest`, if applicable, are added to `other` elements as well.

amples of usage. The first column indicates hyperonym of each synset, displaying its first literal and sense number.

The sense numbers to select are immediately after the emphasised words in the text, as selection boxes. The sense that semyhe offered to the word is pre-selected. User has to select appropriate sense, and after finishing (or leaving the program) save his/her work with appropriate button in the lowest frame.

4. Problems of compatibility

4.1. Part of speech, WSD and syntax

There parts of speech used in EuroWordNet are: noun, proper noun, verb, adjective, adverb. Semyhe looks at Estmorf output only for nouns and verbs. With noun it gets, by default, `_S_ com` and `_S_ prop` — that is substantives and proper nouns. In EWN, numerals (`_N_ card` and `_N_ ord` in Estmorf output) are classified also as nouns. Seems to be a minor bug, but there is a famous example of homonymy in Estonian: ‘viis’ means number ‘five’, and ‘a way to do something’, and ‘melody’. For an English analog, consider the homophony of ‘4’ and ‘for’, for example. By using morphologically disambiguated text, we have already pre-selected one sense (or reduced the possible number of senses) and left the other(s). The same stands for some features, that belong to syntax: verb may be main, auxiliary or modal, by determining the type, we can tell the sense.

4.2. A word about encoding

As Latin alphabet is used to write Estonian, it seems that there should not be a problem with encoding. There are some umlaut letters in Estonian (ä, ö, ü and Ä, Ö, Ü) that rise no problems, since they can be found in many West-European languages and in Latin-1 encoding as well. Some ten years ago there have been some problems with another quite frequent letter ‘õ, Õ’, known as o tilde. It is in Latin-1 now and is OK, but historically there have been problems, as it was not included in so-called ‘extended ASCII character set’ provided by first PC-s running DOS.

There are some really ‘nasty’ letters in Estonian alphabet, s caron and z caron (š, ž, Š, Ž). They are not very frequent, but they figure in important foreign words like ‘žanr’ (genre), ‘dušš’ (shower), or ’garaaž’ (garage), that do not have any synonyms without these ‘horned’ letters. There have been proposals to stop using them and replace them with ‘sh’ and ‘zh’, like in English word ‘bush’, but it can happen in Estonian that syllable boundary—or even word boundary in compounds—is between ‘s’ and ‘h’ like in ‘klaashelmes’ (klaas+helmes, glass bead), so it is not reasonable to use ‘sh’ as ligature. These letters are not contained in Latin-1 character set.

The new standard sets Latin-15 as character set of Estonian, but many applications do not recognise it yet.

The caron letters are in Latin-2 (Windows 1250, Central Europe) encoding, but the places of ‘õ’ and ‘Õ’ are taken by ‘ő’ and ‘Ő’ (o with double acute, used in Hungarian). The bad news is, that our Polaris uses Windows 1250 encoding, and so are the export files. In order to get relevant results about words containing ‘š’ and ‘ž’, we had to convert the EWN export files into Latin-15 before applying semyhe. Still, some XML tools do not recognize Latin-15 encoding, so we must rebuild everything for at least UTF-8 encoding, to get rid of constant converting to and forth.

4.3. Multi-word expressions

There is still a problem with multi-word expressions. Semyhe does not recognise multi-word expressions at present stage, and so they get no sense number, nor display in thesaurus frame (unless they are synonyms of some one-word literal). So lexicographers have to mention the multi-word units separately in the comment field. The problem is more accute with multi-word verbs, as they may consist of words the senses of which by themselves have little, if anything, to do with the meaning of the whole phrase. Fortunately enough, we are going to have a representative list of of Estonian phrasal verbs and idiomatic expressions by Kaalep and Muischnek (2002). Even with the ready-made list the algoritm of founding multi-word semantic units from the text would not be trivial. The same problems that Kaalep and Muischnek met at compiling the database, will haunt us at finding multi-word units from text by semyhe: relevant words may be intervened by other words in the sentence, and we need to meet the ends of lexicon form of word (lemma) and the form that is used in the text. The question of multi-word phrases, verbs in particular, is not a minor one, as there are 1070 two-word phrases in EstWN, 824 of them verb phrases. That makes about 28% of all verb literals[4].

5. Conclusions and future improvements

The tool has turned out to be useable, but there are problems as well, some of them being technical, some theoretical.

We are using morphologically disambiguated text. For semantic analysis, only these nouns and main (or modal) verbs are presented, that are currently in the thesaurus. If there has been made a mistake during morphological disambiguation, a lexicographer using the program can not make any corrections directly, but only make notes about the mistake in the comments field.

Multi-word phrases missing from analysis is a serious drawback, especially in case of verbs. If user does not see the possibility of multi-word phrase in the thesaurus, then it takes him or her much more time to think about this possibility among others. This slows down the process of analysis and is a potential source of errors.

The possibility to see all senses of a word together, in one table, is an advantage that even Polaris does not afford. This gives us direct comparison of senses, that is useful not only for WSD task, but for improvement of the thesaurus as well.

6. References

E. Agirre and G. Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen.

[4] These figures are calculated on EstWN version kb39, as it was used for Senseval-2

J. Dowdall, M. Hess, N. Kahusk, K. Kaljurand, M. Koit, F. Rinaldi, and K. Vider. 2002. Technical terminology as critical resource. To be published in LREC 2002 Proceedings.

H.-J. Kaalep and K. Muischnek. 2002. Using the text corpus to create a comprehensive list of phrasal verbs. To be published in LREC 2002 Proceedings.

H.-J. Kaalep. 1997. An estonian morphological analyser and the impact of a corpus on its development. *Computers and the Humanities*, 31:115–133.

C. Kunze and L. Lemnizer. 2002. Adapting GermaNet for the web. In *Proceedings of 1st International Global Wordnet Conference, January 21–25, 2002*, pages 174–181, Mysore, India. Central Institute of Indian Languages.

M. Louw. 1998. Polaris User's Guide. The EuroWordNet Database Editor. EuroWordNet (LE-4003), Deliverable D023D024.

T. Pavelek and K. Pala. 2002. Visdic — a new tool for wordnet editing. In *Proceedings of 1st International Global Wordnet Conference, January 21–25, 2002*, pages 192–195, Mysore, India. Central Institute of Indian Languages.

T. Puolakainen. 2001. *Eesti keele arvutigrammatika: morfoloogiline ühestamine*. Ph.d. diss., University of Tartu. In Estonian. English title: Computer Grammar of Estonian: Morphological Disambiguation.

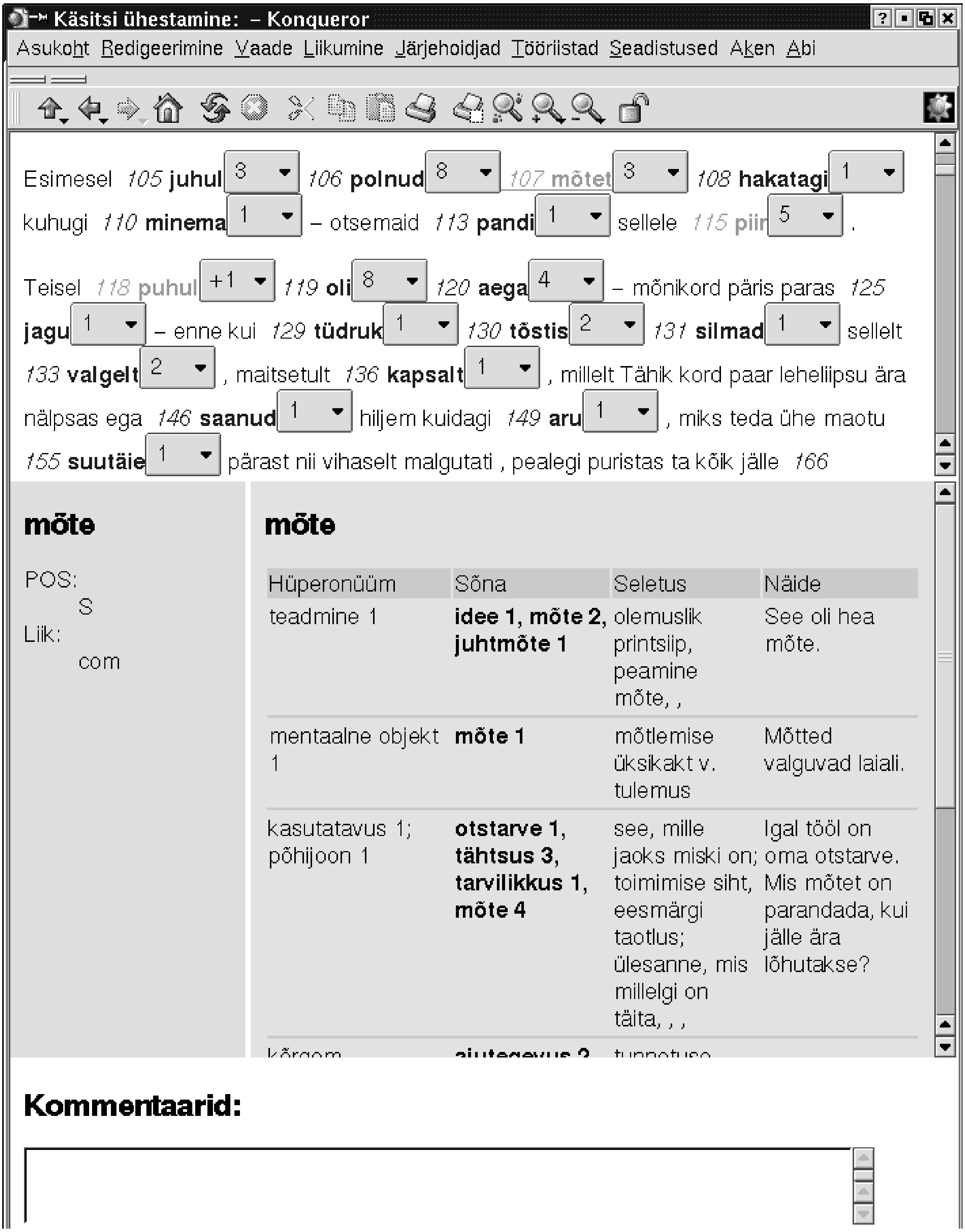
T. Roosmaa, M. Koit, K. Muischnek, K. Müürisep, T. Puolakainen, and H. Uibo. 2001. *Eesti keele formaalne grammatika*. University of Tartu, Tartu, Estonia. In Estonian. English title: The Formal Grammar of Estonian.

P. Smrz. 2002. Storing and retrieving WordNet database (and other structured dictionaries) in XML lexical database management system. In *Proceedings of 1st International Global Wordnet Conference, January 21–25, 2002*, pages 201–206, Mysore, India. Central Institute of Indian Languages.

K. Vider and K. Kaljurand. 2001. Automatic WSD: Does it make sense of Estonian? In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Diasambiguating Systems*, pages 159–162.

```
0 @55718@ WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "job"
      3 SENSE 2
      3 DEFINITION "what you should do for a living"
      3 EXTERNAL_INFO
        4 SOURCE_ID 1
          5 TEXT_KEY "08508615-n"
2 LITERAL "work"
  3 SENSE 1
  3 STATUS "New"
     3 DEFINITION "what you do for a living"
  3 USAGE_LABELS
    4 USAGE_LABEL "sub"
      5 USAGE_LABEL_VALUE "Medicine"
  3 FEATURES
    4 FEATURE "connotation"
      5 FEATURE_VALUE "figurative"
   /---/
-------------------------------------------------------------------
<?xml version="1.0"?>
<THESAURUS>
<WM ID="55718">
  <POS VALUE="n"/>
  <VARIANTS>
    <LITERAL VALUE="job">
      <SENSE VALUE="2"/>
      <DEFINITION VALUE="what you should do for a living"/>
      <EI>
        <SI VALUE="1">
          <TK VALUE="08508615-n"/>
        </SI>
      </EI>
    </LITERAL>
    <LITERAL VALUE="work">
      <SENSE VALUE="1"/>
      <STATUS VALUE="New"/>
      <DEFINITION VALUE="what you do for a living"/>
      <USLS>
        <UL VALUE="sub">
          <ULV VALUE="Medicine"/>
        </UL>
      </USLS>
      <FEATURES>
        <FEATURE VALUE="connotation">
          <FV VALUE="figurative"/>
        </FEATURE>
      </FEATURES>
    /---/
  </VARIANTS>
</WM>
</THESAURUS>
```

Figure 4: An extraction from EWN import-export format (upper) translated into XML format (lower)


Käsitsi ühestamine: – Konqueror
Asukoht Redigeerimine Vaade Liikumine Järjehoidjad Tööriistad Seadistused Aken Abi

Esimesel *105* **juhul** 3 ▾ *106* **polnud** 8 ▾ *107* **mõtet** 3 ▾ *108* **hakatagi** 1 ▾ kuhugi *110* **minema** 1 ▾ – otsemaid *113* **pandi** 1 ▾ sellele *115* **piir** 5 ▾ .

Teisel *118* **puhul** +1 ▾ *119* **oli** 8 ▾ *120* **aega** 4 ▾ – mõnikord päris paras *125* **jagu** 1 ▾ – enne kui *129* **tüdruk** 1 ▾ *130* **tõstis** 2 ▾ *131* **silmad** 1 ▾ sellelt *133* **valgelt** 2 ▾ , maitsetult *136* **kapsalt** 1 ▾ , millelt Tähik kord paar leheliipsu ära nälpsas ega *146* **saanud** 1 ▾ hiljem kuidagi *149* **aru** 1 ▾ , miks teda ühe maotu *155* **suutäie** 1 ▾ pärast nii vihaselt malgutati , pealegi puristas ta kõik jälle *166*

mõte

POS:
S
Liik:
com

mõte

Hüperonüüm	Sõna	Seletus	Näide
teadmine 1	**idee 1, mõte 2, juhtmõte 1**	olemuslik printsiip, peamine mõte, ,	See oli hea mõte.
mentaalne objekt 1	**mõte 1**	mõtlemise üksikakt v. tulemus	Mõtted valguvad laiali.
kasutatavus 1; põhijoon 1	**otstarve 1, tähtsus 3, tarvilikkus 1, mõte 4**	see, mille jaoks miski on; toimimise siht, eesmärgi taotlus; ülesanne, mis millelgi on täita, , ,	Igal tööl on oma otstarve. Mis mõtet on parandada, kui jälle ära lõhutakse?
kõrgem	**ajutegevus 2**	tunnetuse	

Kommentaarid:

Figure 5: The user interface of the lexicographer’s tool as seen in Konqueror browser. In upper frame, there is current text; in left part of middle frame (gray background), there is some morphological information (part of speech and class), in right part of middle frame there is semantic information from Estonian WordNet presented in a table; the lowest frame is for lexicographer’s comments.

Approaches to Checking Subsumption in GermaNet

Piklu Gupta

Fraunhofer Integrated Publication and Information Systems Institute
Dolivostr. 15
D-64283 Darmstadt, Germany
gupta@ipsi.fraunhofer.de

Abstract

The paper describes different approaches for checking the subsumption relation in GermaNet using database queries and subsequent manual analysis. The work was carried out in an object-oriented tool environment hosting the GermaNet data. Finally there is a brief note comparing GermaNet coverage with that of Duden dictionaries.

1. Introduction

The context of the work presented here was a study for Bibliographisches Institut & F.A. Brockhaus (BIFAB), publishers of Duden dictionaries; the main purpose of the study was to subject GermaNet (Hamp and Feldweg, 1997; Kunze, 2000) to close scrutiny by examining semantic relations in terms of consistency and subsequently to compare coverage of GermaNet with Duden dictionaries. The relations we focused on were the generic hierarchical relation expressed by hyponymy/hyperonymy and its synonymous variant for verbs (troponymy/troponymyOf), since this is the fundamental relation in GermaNet (Kunze, 1999). The tool hosting GermaNet for this work was the TerminologyFramework system, briefly described below in subsection 2.1. Various approaches to investigating the subsumption relation were adopted:

- formal consistency checks using database queries and manual analysis of results
- manual inspection of the non-overlapping parts of a semantic field and the corresponding concept hierarchy in GermaNet
- manual inspection of subsumption links reachable from a 10% sample of the denotation strings of GermaNet which also belong to the single volume Duden dictionary (DUDEN, 2000b) as lexical entries.
- top concepts were identified and analysed.

The starting point for the manual tests was rigid or strict subsumption: concept A is subsumed by concept B iff all instances of A are also instances of B.

Duden made their dictionary material available to us in machine-readable format thus enabling us to also compare the coverage of GermaNet with both the 10 volume (DUDEN, 2000a) and the single volume Duden dictionaries.

2. Formal Consistency Checks with Queries in Terminology Framework

2.1. Terminology Framework

TerminologyFramework (henceforth TFw) is a general purpose tool for representation and maintenance of thesaurus-like structures, ranging from conventional thesauri to the published CyC upper ontology or lexical databases such as WordNet (Fischer, 1998). GermaNet was imported into a TFw application, using an identical schema previously developed for investigating WordNet (Fischer, 1997). This import generated an object-oriented representation of GermaNet including persistent storage. Its contents could be inspected with tools including a frame to slot to value list view and graphic view (described in Möhr and Rostek (1993)) and investigated by means of database queries. The import turns every synset into an object (known in TFw as a **concept**) and the synset elements are represented as **terms**, which are denotation objects with disambiguated denotation strings. One of the advantages of TFw is that it allows for computable relations such as the transitive closure of the subsumption relation.

2.2. Formal Consistency Checks

A broad range of formal checks for redundancy and consistency in WordNet had already been devised and described by Fischer (1997). We restricted ourselves to consistency checks with respect to the subsumption relation in GermaNet. Fischer's investigation employed three distinct queries relevant to this relation:

1. Are there opposed concepts where one subsumes the other?
2. Are there opposed concepts which have a common subconcept?
3. Are there examples where the commutativity of subsumption and opposedness does not hold?

We understand subsumption not only as a relation that holds directly but also indirectly between concepts (as a result of the transitivity of this relation), which means that these questions presuppose the availability of the transitive closure of hyponymy/hyperonymy in GermaNet. The 'opposed' relation is defined thus: two concepts are opposed (or synonymously 'antosemous') if at least two of their terms are antonyms. Therefore a further computable semantic (concept-concept) relation is induced from a lexical (term-term) relation and both computable relations are prerequisites for the check. If we consider the third query, we need to explain what is meant by commutativity of subsumption and opposedness. Fischer (1997) defines this as follows:

> For each concept *c:* If *antosem(c)* is not empty, then the equation *hypernym(antosem(c)) = antosem(hypernym(c))* or set inclusion in one direction or the other should hold.

All three rules may be justified by a concept model with feature inheritance, assuming that opposed concepts necessarily have some kind of contradictory feature which must not be inherited simultaneously by a more specific concept, otherwise this would lead to an oxymoron (e.g. bittersweet). We use this example, however, to illustrate that it is by no means impossible for language to creatively violate this logical inheritance rule. These three questions are therefore best seen as a heuristic to detect on the one hand cases which entail errors and on the other hand cases which invalidate the generality of the rule.

We did not consider the last of the three questions concerning commutativity, but concentrated instead on the first two. The retrieval results are discussed below in subsections 2.3. and 2.4.

2.3. Does GermaNet contain opposed concepts where one subsumes the other?

This query posed to the classes of verb and adjective concepts returned no hits, but when posed to the class of noun concepts it returned three noun concept pairs, illustrated by the three figures below:

- *Ziegenbock* (male goat) and *Ziege(2)*[1] (goat, in the generic rather than female sense),
- *Subjekt(2)* (subject in the sense of a living being) and *Objekt(2)* (object in the sense of living being)
- *Titelverteidiger* (title holder) and *Herausforderer* (challenger)

Figure 1 illustrates the case of *Ziegenbock*. Here we see that the antonymy relation has been falsely assigned between the generic and the male form; there should be a link showing antonymy stretching from left to right in the figure, that is from the term *Ziegenbock* to the term *Ziege* of the concept *Ziege* in its female sense. This case appears to be the result of an incorrectly assigned pointer due to homographs; we can only speculate as to whether inappropriate tools or limited views used in linking concepts by the lexicographer are the source of the error here.

Figure 2 illustrates the case of *Subjekt (2)* and *Objekt(2)*. The opposed relation between *Subjekt(2)* and *Objekt (2)* induced by antonymy is clearly false. We suggest that another pair of concepts, ‘namesakes’ to the given pair – *Subjekt (1)* and *Objekt (1)*, both in the grammatical sense, should be linked as opposed concepts. The ‘namesakes’ relation is a computable TFw relation which links concepts with homographic denotation strings.

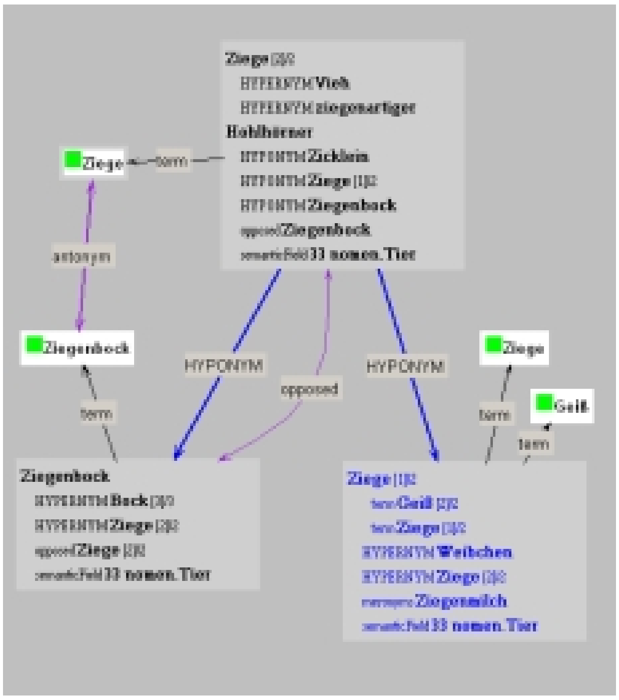


Figure 1: Faulty antonymy target

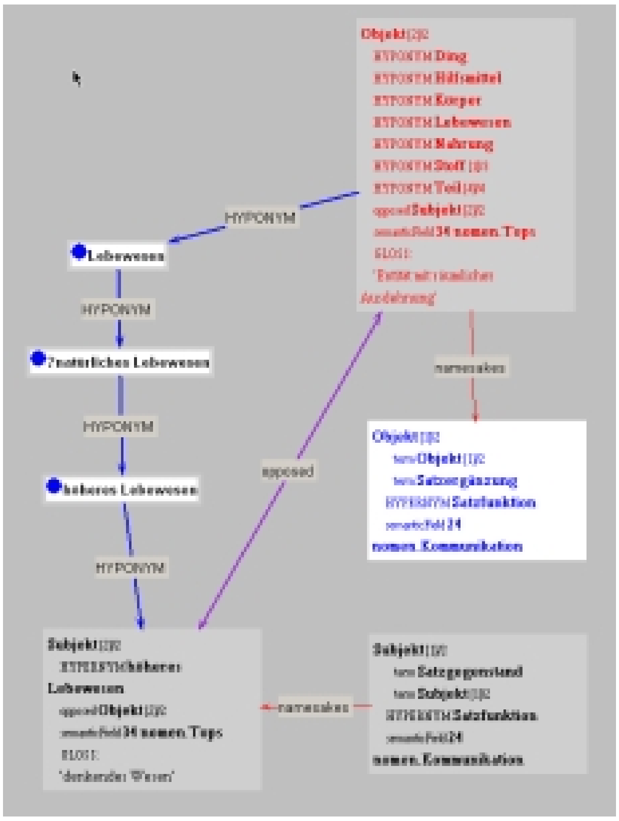


Figure 2: Faulty antonymy pair

The case of *Titelverteidiger* (title holder) and *Herausforderer* (challenger), illustrated in Figure 3 below, leads to a different diagnosis. We maintain that the hyponym link between *Herausforderer* and *Champion* (champion) is incorrect, since not every champion is a challenger. [2]

[1] The number after the word denotes a homograph counter generated by TFw; the figures also show the number of homographs for each respective homographic string, separated by a ‘/’.

[2] Note that an antonym link is missing between the gender-inclusive forms *HerausforderIn* and *TitelverteidigerIn*.

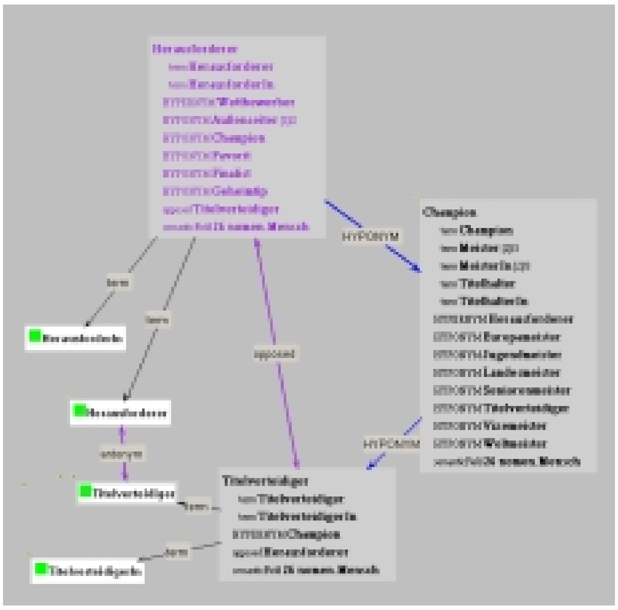


Figure 3: Faulty hyponym link between *Herausforderer* and *Champion*

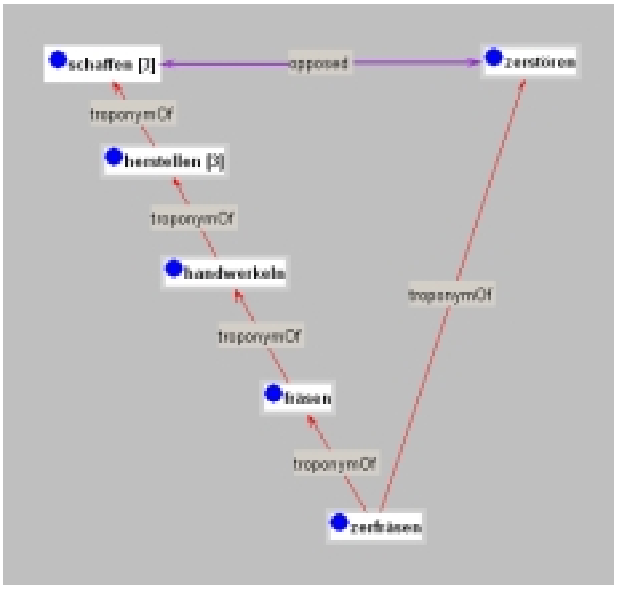


Figure 4: Strict versus defeasible hyperonymy

2.4. Does GermaNet contain opposed concepts with a common hyponym?

The query posed to the class of verb concepts returned two verb concept pairs, illustrated by the two figures below:

- *schaffen (3)* (in the sense of to create) and *zerstören* (to destroy); their common troponyms are *zersägen* (to saw up), *zerkochen* (to overcook or cook to a pulp), and *zerfräsen* (to mill to pieces)

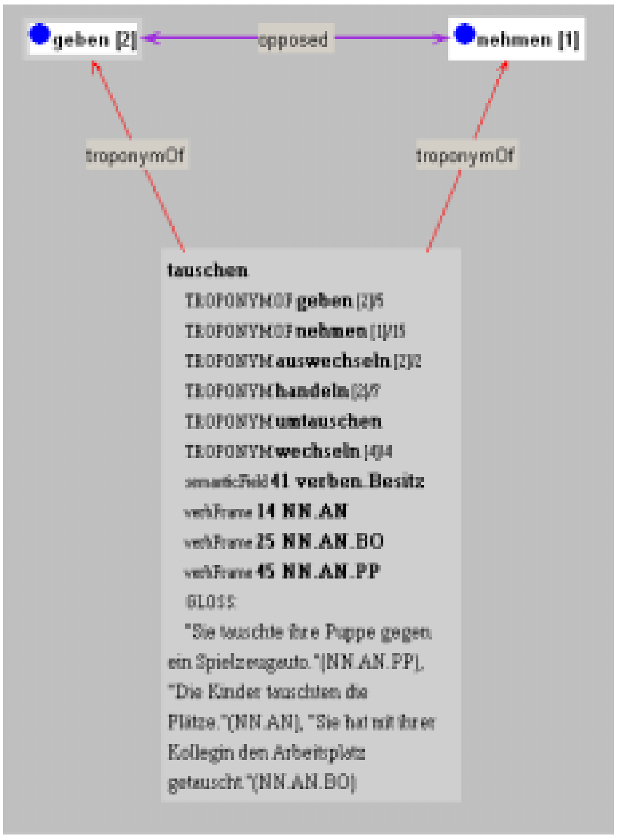
- *nehmen (1)* (in the sense of to take something) and *geben (2)* (in the sense of to give something); there are 8 common troponyms including e.g. *tauschen* (to exchange something for another thing) and *dealen* (in the sense of dealing e.g. drugs).

This query returned results which are not indicative of incorrect pointer assignment but rather raise non-trivial questions about the nature of the subsumption relation or the antonymy relation in GermaNet.

The figure shows that *zerfräsen* is simultaneously a hyponym of verb concepts denoting creation and destruction. At first sight this seems counterintuitive. How can we account for this phenomenon? The hyponymy relation of *zerfräsen* and *zerstören* is obviously correct and is a rigid subsumption link. Looking at the left hand side of the figure, we check the link from *zerfräsen* to *fräsen*. This we deem to be acceptable as a rigid subsumption link, if we *fräsen* means to use a milling tool or mould in its neutral sense irrespective of its creative or destructive effect. If, however, we proceed from that concept node upwards to *schaffen (3)* we leave the neutral sense of *fräsen* and adopt a sense in which a creative or non-destructive use of the tool is implicit. It therefore follows that we have given the concept node *fräsen* two different meanings, and therefore according to the general WordNet philosophy we should split the node into 3: *fräsen(neutral)* , *fräsen (constructive)* and finally *fräsen (destructive)*, which already exists as *zerfräsen*.

Another possible remedy is to differentiate between strict and defeasible (non-strict) subsumption; the link between *zerfräsen* and *fräsen* would be strict whereas the link between *fräsen* and its direct superordinate *handwerken* or its indirect superordinate *schaffen (3)* is non-strict, i.e. in most cases the use of a milling tool or mould is constructive. Introducing a new subsumption relation type to the WordNet software, however, is likely to be difficult in contrast to TFw. This would entail checking all subsumption links for their type. Note that we cannot assume transitivity for the concatenation of strict and non-strict subsumption links.

A radically different diagnosis and remedy spring to mind when considering the case of Figure 5. At first sight the constellation appears to be acceptable, thus disproving the general validity of the rule. Our intuition may tell us that *tauschen* implies simultaneous acts of giving and taking and thus even the conjunction of the superordinates *nehmen* and *geben* seems plausible. On closer inspection, however, we see that a *tauschen* act implies the taking of **one item** in exchange for **another**, which means that the act of exchange consists of two simultaneous (or more probably) consecutive acts of giving X and taking Y where X and Y are not identical. The opposition of the concepts ‘giving’ and ‘taking’, however, obviously implies that the object of both is the same otherwise there would be no opposition. For example, teaching linguistics is not the ‘opposite’ of learning mathematics. What does the antonym or opposed link actually mean? (cf. Woods (1991, pp. 54ff)) If it means every act of giving is opposed to every act of taking, in the same way as every sweet object is opposed to every savoury object then the opposed link is faulty. If it means that for every giving act there exists a taking act which is opposed, then the rule implicit in the query does not have general validity! This demonstrates the inconsistent use of the antonym/opposed link. Instead of the troponymOf links between *tauschen* and *nehmen* and *tauschen* and *geben* we propose a pair of ‘entails’ links, which would show that exchanging entails both giving and taking.



tauschen
TROPONYMOF geben [2]/5
TROPONYMOF nehmen [1]/15
TROPONYM auswechseln [2]/2
TROPONYM handeln [3]/7
TROPONYM umtauschen
TROPONYM wechseln [4]/4
semanticField 41 verben.Besitz
verbFrame 14 NN.AN
verbFrame 25 NN.AN.BO
verbFrame 45 NN.AN.PP
GLOSS:
"Sie tauschte ihre Puppe gegen ein Spielzeugauto."(NN.AN.PP), "Die Kinder tauschten die Plätze."(NN.AN), "Sie hat mit ihrer Kollegin den Arbeitsplatz getauscht."(NN.AN.BO)

Figure 5: *Geben (2)* and *nehmen (2)* only opposed with a common object

Posing the query to the class of adjective concepts returned one concept pair, *farbig (2)* (in the non-racial sense of coloured) and *farblos* (colourless) with the common hyponym *falb* (dun, as applied to horses). This constellation contains a highly questionable hyponym link between *fahl* (pale) and *farblos* (colourless).

Posing the same query to the class of noun concepts also returned a single concept pair, *Vermögen* (property) and the non-lexicalised concept *?negativer Besitz* (negative ownership). In this case two highly questionable hyponym links exist, on the one hand between *Zins* (interest) and *Vermögen* or *Finanzen* (finances) and on the other between *Verzugszins* (interest payable on arrears) and *?negativer Besitz*.

In concluding this section, we note that retrieval results for both kinds of questions did not invalidate the implicit heuristic rules.

3. Semantic fields and hyponymy

According to GermaNet documentation (http://www.sfs.nphil.uni-tuebingen.de/lsd) , the division of GermaNet into semantic fields served an organisational purpose in that a field corresponds to a data file for editing by lexicographers. It was nonetheless interesting to investigate to what extent the semantic fields did in fact contain the expected content and to this end we looked at the hyponymy relation in cases where an available top concept corresponded to a semantic field label. Wherever this proved to be the case, we would expect all hyponyms to be members of that semantic field. Those elements **not** in the intersection of both sets demanded closer inspection, for instance with regard to which hierarchy they should actually belong to. This is another fruitful method for delimitation of the set of hyponym links to be checked manually.

Among others, the semantic field *nomen.Tier* (noun animal) was examined in tandem with the concept *Tier* (animal) and queries led us to obtain the following results:

- the concept *Tier* has 2049 hyponyms.
- the semantic field contains 2086 elements.
- only one concept *Pute* (turkey in its food sense) is an indirect hyponym of *Tier* but not a member of the semantic field *nomen.Tier* and instead belongs to the semantic field *nomen.Nahrung* (noun food). Here we have a clear-cut case of ‘animal grinding’ (Briscoe et al., 1995), in which a count noun (animal) becomes a mass noun (food). It might therefore be useful to assign a different kind of link which is applicable to the grinding operation.
- 38 concepts are in the semantic field but are not hyponyms of *Tier*; almost half of these consist of mythical beasts. The remainder include borderline cases such as single-celled beings, bacteria and microorganisms. If we maintain that mythical beasts such as *Einhorn* (unicorn) are animals irrespective of their real existence, then they should by rights also be hyponyms of *Tier*. There are also concepts such as *Männchen* (male animal) and *Weibchen* (female animal) which should properly be classed as animals.

4. Missing links

Formally speaking, top concepts are those which have no superordinates. GermaNet contains 500 such formal top noun concepts, but it should be noted, however, that of these 500 concepts only 125 are what we would term genuine top concepts in that they also have hyponyms, the remaining 375 are therefore isolated having neither superordinates nor hyponyms. For verbs there are 125 genuine tops and 94 isolated concepts and for adjectives there are 34 genuine tops and 246 isolated concepts. This points towards the transitional status of these concepts – GermaNet is, after all, a work in progress. The large number of remaining top concepts for nouns and verbs in particular is therefore arguably due to missing structure at this highest level. For example, a number of the genuine tops should either be hyponyms of other tops or hyponyms of new, more general concepts – *Wurstware* (sausages) is a top and is not a hyponym of *Nahrung* (food) as would be expected and tops such as *Arbeitszeit* (working hours), *arbeitsfreie Zeit* (leisure time) are not linked to the possible concept of time interval. Another approach to finding missing antonym links is to run the third query mentioned above in subsection 2.2. and discussed for WordNet in Fischer et al. (1996, p. 253) and Fischer (1997, p. 28).

5. Manual evaluation of generic links in a sample of GermaNet

This section summarises results of an investigation of a sample of GermaNet with regard to the correctness of the hyponymy relation. The basis of the sample was a list of GermaNet synset elements which also appear in the single volume Duden dictionary as lexicon entries. Starting with the ninth list entry and subsequently every tenth entry was extracted from a list of adjectives, nouns and verbs thus providing us with a 10% sample of GermaNet. A total of 3511 hyperonym links were examined and classical tests for strict hyponymy were applied. We distinguished between correctly assigned, doubtful [3] and incorrectly assigned hyperonymy. Results were as follows:

- out of 519 verb denotation strings we derived 914 hyperonym links , 89% were deemed to have correctly assigned hyperonymy, 4% were doubtful and 7% were incorrect,
- out of 396 adjective denotation strings we derived 489 hyperonym links, 92% were correct, 2.5% were doubtful and 5% were incorrect,
- out of 1664 noun denotation strings we derived 2108 hyperonym links, 96.6% were correct, 1.2% were doubtful and 2.2% were incorrect,
- of a total of 2579 denotation strings for all 3 GermaNet word classes we derived 3511 hyperonym links, 94.1% were correct, 2.1% were doubtful and 2.1% were incorrect.

Some of the commonest errors were mistaken assignment of hyponym/troponym where a merge of concepts would be more appropriate because their terms are stylistic variants and therefore synonyms. For instance, the stylistic variants *pennen*, *knacken* and *ratzen* (to kip; colloquial for to sleep) are deemed to be troponyms of *schlafen* (to sleep) rather than as what Cruse (1986) regards as ‘cognitive synonyms’. The assignment of hyponymy seemed on occasions to be based on morphological factors rather than semantic ones (e.g. *Fahrgast* (passenger) as a hyponym of *Gast* (guest).

6. Coverage of GermaNet compared with Duden dictionaries

It is a truism that both the single volume and 10 volume Duden dictionaries have wider coverage than GermaNet, with around 100,000 entries and 200,000 entries respectively so it is arguably more interesting to look at what is to be found in GermaNet but not in single or multi-volume reference works rather than to simply enumerate what is in the dictionary but not in GermaNet. GermaNet contained [4] a total of 41359 entry strings, of which 25798 appear in both GermaNet and the single volume Duden. A total of 15561 entry strings were to be found in GermaNet but not in the single volume Duden. 28862 entry strings appeared in both GermaNet and the 10 volume Duden and 12497 entry strings were present in GermaNet and not in the 10 volume Duden.

A number of groups in GermaNet and in neither of the Duden dictionaries can be identified as follows:

- gender-neutral terms denoting roles (e.g. *AntifaschistIn* (anti-fascist))
- very specific specialised language (e.g. terms from a biological taxonomy)
- selected compounds; compounding is highly productive in German and therefore criteria for their selection and inclusion are dependent on e.g. frequency, corpus evidence
- orthographic variants
- misspellings

GermaNet contains 1869 gender-neutral terms denoting roles which are not present in the form with an upper case ‘I’ in either the single or 10 volume Duden, but feminine forms are to be found if the upper case I is eliminated by a normalisation to lower case letters. GermaNet appears to contain an exhaustive biological taxonomy (despite claims for inclusion on the basis of corpus frequency), so on inspection of the 2049 hyponyms of *Tier* (animal) and the 189 hyponyms of *Pflanze* (plant), 1043 animal hyponyms and 1119 plant hyponyms are present that are not to be found in the 10 volume Duden. The difference between what is present in GermaNet and in dictionaries raises important questions for lexicographers – for instance, which criteria should be employed for inclusion of compounds, which can in any case never be completely covered due to the productivity of compounding. Also, how subjective frequency decisions made by lexicographers are and to what extent the use of balanced corpora can contribute to lexicography.

7. Acknowledgements

I am grateful to Dietrich Fischer for his insights and encouragement in the writing of this paper. Constructive discussion with him proved invaluable. I would also like to thank Lothar Rostek for initiating the study in the first place and for playing a major role in setting up the working environment and the scenario.

8. References

Ted Briscoe, Ann Copestake, and Alex Lascarides. 1995. Blocking. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational Lexical Semantics*, pages 273–301. Cambridge University Press, Cambridge.

D A Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge.

DUDEN. 2000a. *Das große Wörterbuch der deutschen Sprache*. Bibliographisches Institut & F.A. Brockhaus A.G., Mannheim.

[3] Some of the links that we deemed in this analysis to be merely doubtful (such as the link between *tauschen* (to exchange) and the *geben* (to give) and *nehmen* (to take) pair) were deemed incorrect after the formal checks described in subsection 2.4..

[4] We used version 3.0, current as of 22.01.01

DUDEN. 2000b. *Der Duden, 12 Bde., Bd.1, Duden Die deutsche Rechtschreibung, neue Rechtschreibung*. Bibliographisches Institut & F.A. Brockhaus A.G., Mannheim.

Dietrich H. Fischer, Wiebke Möhr, and Lothar Rostek. 1996. A modular, object-oriented and generic approach for building terminology maintenance systems. In Christian Galinski and Klaus-Dirk Schmitz, editors, *TKE '96:Terminology and Knowledge Engineering*, pages 245–258, Frankfurt a.M. INDEKS Verlag.

Dietrich H Fischer. 1997. Formal redundancy and consistency checking rules for the lexical database WordNet 1.5. In Vossen et al. (Vossen et al., 1997), pages 22–31.

Dietrich H Fischer. 1998. From Thesauri towards Ontologies? In Widad Mustafa el Hadi, Jacques Maniez, and Steven A. Pollitt, editors, *Structures and Relations in Knowledge Organisation:Proceedings of the Fifth International ISKO Conference*, volume 6, pages 18–30, Lille, France. Ergon Verlag.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In Vossen et al. (Vossen et al., 1997), pages 9–15.

Claudia Kunze. 1999. Semantics of Verbs within GermaNet and EuroWordNet. In V. Kordoni, editor, *Proceedings of the ESSLLI-99 Workshop on 'Lexical Semantics and Linking in Constraint-Based Theories'*, pages 189–200, Utrecht.

Claudia Kunze. 2000. Extension and use of GermaNet, a lexical-semantic database. In *Proceedings of LREC 2000 2nd International Conference on Language Resources and Evaluation*, Athens.

Wiebke Möhr and Lothar Rostek. 1993. TEDI: An Object-Oriented Terminology Editor. In Klaus-Dirk Schmitz, editor, *TKE '93:Terminology and Knowledge Engineering*, pages 363–374, Frankfurt a.M.

Piek Vossen, Gert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors. 1997. *Automatic Information Extraction and Building of Lexical Semantic Resources*. Association for Computational Linguistics, 12 July 1997.

William A. Woods. 1991. Understanding subsumption and taxonomy: A framework for progress. In John Sowa, editor, *Principles of Semantic Networks: Explorations in the Representation of Knowledge*, pages 45–94. Morgan Kaufmann, San Mateo, CA.

Evaluation of GermanNet: Problems Using GermaNet for Automatic Word Sense Disambiguation

Jahn-Takeshi Saito, Joachim Wagner, Graham Katz, Philip Reuter, Michael Burke, Sabine Reinhard

Institute for Cognitive Science
University of Osnabrück
49074 Osnabrück
Germany

Abstract

WordNets such as GermaNet have been frequently used as an inventory of word-senses for word-sense disambiguation tasks. In the work described here we evaluate the adequacy of GermaNet for this task. That is we attempt to determine the degree to which GermaNet provides an adequate inventory of senses for word-sense annotation of running text. Our findings were on the whole very encouraging. GermaNet provides an appropriate sense for 83 % of the content words in our texts. More interestingly, an error analysis showed that simple morphological processing could significantly improve coverage.

1. Introduction

The use of WordNet for sense tagging of English is by now an established research program (Miller, et.al 1994; Resnik 1998; Landes, Leacock & Tengi 1998). With the advent of WordNet-style lexical resources for languages other than English (Bloksma, Díez-Orzas & Vossen 1996) the application of these resources, to sense-tagging for these languages is a natural evolution. A number of questions arise in this context, however. While the original WordNet has been used with success for English, there is no guarantee that this experience generalizes to other WordNets for other languages. Both the language itself and the particular WordNet developed for it may present problems that were not present in the WordNet/English case. Our goal here was to evaluate how useful GermaNet is as a resource for word sense tagging for German.

Our task, then, was to annotate a corpus of German text using GermaNet and to determine how close to the ideal of providing an appropriate sense tag for all content words in the corpus GermaNet is. This is of interest both as an evaluation of GermaNet itself, and also because German and English differ in ways that, *a priori*, might indicate that German would be a difficult language to sense-tag (Hamp & Feldweg 1997). German has, for example, highly productive word-formation processes and a rich derivational morphology.

In form, however, our work was very similar to that done for English by Landes, Leacock & Tengi (1998) in that we simply set out to manually disambiguated words in a corpus, tagging each appearance of a content word in the corpus. In contrast to their work, we developed our own (German language) corpus and used GermaNet as our repository of word senses. Additionally, work was separate from the development of GermaNet and we did not have contact with GermaNet lexicographers.

2. GermaNet

GermaNet is a lexical-semantic net based on the WordNet example (Kunze & Wagner 1999a). It is intended to cover the basic vocabulary of German. Although GermaNet relies on the design principles and shares the same database structure as the Princeton WordNet (Miller 1990), it is build from scratch and features some modifications. In contrast to WordNet, GermaNet includes non-lexicalized *artificial* concepts to fill lexical gaps (e.g. to provide the missing antonym for *thirsty*) and to avoid unjustified co-hyponomy. Additionally, cross-classification of concepts, which is seldom used in WordNet, is an essential feature of GermaNet, and regular polysemy is integrated via a special relation between synsets. There are also some particular differences with respect to the way parts of speech are handled. Adjectives in GermaNet, for example, are hierarchically structured (in contrast to a clustering approach in WordNet). It wasn't clear that any of these differences affected the usefulness of GermaNet for sense-tagging, however.

More important was GermaNets coverage. Although GermaNet is comparable in size to WordNet, it is significantly smaller, as indicated in Table 1.

	GermaNet	WordNet 1.7
Noun	27824	74488
Verb	8810	12754
Adjective	5141	18523
Adverb	2	3612
Total:	41777	109377

Table 1. GermaNet vs. WordNet

Although GermaNet has been integrated into EuroWordNet (Kunze & Wagner 1999b), the version we used for our research was the stand-alone GermaNet.

3. The annotation task

As a preliminary to the development of an automatic sense tagger for German we hand-tagged eleven small German texts. We used these hand tagged texts to evaluate the applicability of GermaNet to large-scale sense tagging applications. The procedure we used for annotation was fairly straightforward. We automatically lemmatized the words and tagged them for part of speech using the Stuttgart TreeTagger. To actually carry out sense tagging, we developed a software tool for presenting words in texts along with their GermaNet synsets, which was used by five annotators to annotate the texts. The texts were annotated on a word-by-word basis, with each token that had been tagged either as a verb, a noun or an adjective presented for word-sense tagging. For words that could not be annotated with GermaNet synsets, the problem that the word appeared to pose was noted by the annotator, if one was apparent. These error-annotations were used to classify the types of words that presented difficulties for sense-tagging using GermaNet synsets.

3.1. Corpus preparation

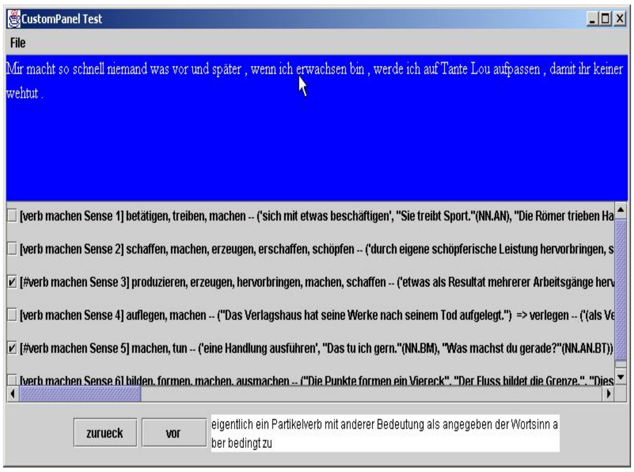
As there is not yet a standard representative German corpus, we choose to develop our own corpus. The corpus consisted of eight short excerpts from novels for children and young people and three articles taken from German newspapers. The total number of words in our corpus was 5625 and the individual subcorpora varied in size from 257 to 1021 words.

The entire corpus was both lemmatized and tagged for part of speech by the IMS TreeTagger (Schmid 1994). These lemmata were then used to automatically compile a list of GermaNet synsets for each token in the corpus that appeared in GermaNet. For each lemma, the complete set of GermaNet synsets associated with the lemma by GermaNet was stored alongside the lemma. The POS information was **not** used in this step for filtering, so as to exclude this as a source of error. As indicated in Table 2, GermaNet assigned a synset to more than 90% of the content words (noun, verb or adjective tagged words) in the texts. Strikingly, the percentage of content words not assigned an appropriate synset by GermaNet is lower for the newspaper corpora (about 80%) then for the childrens fiction corpora (about 85%).

3.2. Corpus Annotation

For purposes of annotation, the eight short-novel corpora were split up randomly into 24 pieces which were recombined into equal-sized subcorpora and distributed among our five annotators. The pieces were systematically permutated in order to minimize the influence of inter annotator differences. After annotation was complete the pieces were reordered, so that statistics could be obtained on a per corpus basis. At a later stage the newspaper subcorpora News 1, News 2, and News 3 were annotated. Although the annotation procedure was the same, these subcorpora were annotated by a single annotator.

The actual annotation was carried out as follows. The five annotators – all native speakers of German – were provided with a software tool and a set of files to be tagged. The software tool (see fig. 1) presented the annotator with each occurrence of a lemma for which GermaNet provided synsets.



Mir macht so schnell niemand was vor und später , wenn ich erwachsen bin , werde ich auf Tante Lou aufpassen , damit ihr keiner wehtut .

☐ [verb machen Sense 1] betätigen, treiben, machen -- ('sich mit etwas beschäftigen', "Sie treibt Sport."(NN.AN), "Die Römer trieben Ha

☐ [verb machen Sense 2] schaffen, machen, erzeugen, erschaffen, schöpfen -- ('durch eigene schöpferische Leistung hervorbringen, s

☑ [#verb machen Sense 3] produzieren, erzeugen, hervorbringen, machen, schaffen -- ('etwas als Resultat mehrerer Arbeitsgänge herv

☐ [verb machen Sense 4] auflegen, machen -- ("Das Verlagshaus hat seine Werke nach seinem Tod aufgelegt.") => verlegen -- ('(als Ve

☑ [#verb machen Sense 5] machen, tun -- ('eine Handlung ausführen', "Das tu ich gern."(NN.BM), "Was machst du gerade?"(NN.AN.BT))

☐ [verb machen Sense 6] bilden, formen, machen, ausmachen -- ("Die Punkte formen ein Viereck", "Der Fluss bildet die Grenze.", "Dies

eigentlich ein Partikelverb mit anderer Bedeutung als angegeben der Wortsinn a
ber bedingt zu

Figure 1. The TAZAN annotation tool

The annotator task was to mark the appropriate synset, if there was one. In addition to the textual context the word appeared in, i.e. the sentence, annotators were shown the set of synsets for the lemma and the basic characterization provided by GermaNet for these synsets. These contained brief descriptions of the synset, examples of typical uses of that sense of the word and an indication of where the synset was located in the GermaNet hierarchy. For verbs the syntactic frame associated with the sense was also indicated. The synsets were presented to the annotator grouped by POS. In choosing a synset, annotators also implicitly indicated what they took to be the correct POS for the word in contexts.

For the lemma *essen*, for example, the following information was presented, with three noun senses and one verbal sense.

[nomen essen Sense 1] Essen, Mahl, Mahlzeit --
 ('Einnahme von Speisen')
[nomen essen Sense 2] Gericht, Speise, Essen --
 ('Speise, die für eine Mahlzeit zubereitet ist') =>

Corpus	Word Tokens	Content words	Synset Assigned	Marked	Marked (of Assigned)
Fiction	4330	1770	1658 (93.7%)	1497 (84.6%)	90.3%
Newspaper 1	257	143	129 (90.2%)	124 (86.7%)	86.7%
Newspaper 2	474	206	179 (86.9%)	161 (78.2%)	89.9%
Newspaper 3	564	270	233 (86.3%)	205 (75.9%)	76.3%

Table 2: Quantitative Characterization of the Corpora and Annotation Results

Nahrung, Nahrungsmittel, Lebensmittel, Esswaren, Eßwaren*o, Essen, Speisen
[nomen essen Sense 3] Nahrung, Nahrungsmittel, Lebensmittel, Esswaren, Eßwaren*o, Essen, Speisen => Objekt -- ('Entität mit räumlicher Ausdehnung')
[verb essen Sense 1] essen, futtern*s, nehmen -- ('etwas zu sich nehmen', "Er isst kein Fleisch."(NN.AN), "Er futtert wie ein Scheunendrescher."(NN.BR), "Sie nimmt viel Flüssigkeit zu sich."(NN.AN.PP), "Die Kinder futtern fleißig Schokolade."(NN.AN.BM) "Heute abend werde ich warm essen."(NN.BM)) => verzehren -- ('Ein Lebensmittel essen oder trinken, Perspektive auf Lebensmittel', "Auf der Weihnachstfeier haben die Mitarbeiter zehn Kilo Fleisch verzehrt.", "Sie verzehrte ihr Gemüse ohne Appetit.")

The annotators were also encouraged to use the GermaNet browser to locate additional information about a synset if a decision was difficult.

To annotate, the annotator simply selected (via check box) the appropriate sense(s) for the word as used in the context presented. They were able to move freely forwards and backwards through the corpus and to change their choice of synset at any time. The task was not an easy one. To fully annotate even one of our 24 small subcorpora took our annotators approximately an hour of annotation time. Typically, however, our annotators divided up the task into a number of sessions.

Note that annotators were instructed to mark all synsets considered appropriate. That means that the annotator could mark more than one of the senses GermaNet assigned to the word or reject all of them. This means that words which were not assigned at least one GermaNet synset were not presented for tagging at all. As indicated in the sixth column of Table 2, this was typically around 10% of the content words.

3.3. Results of annotation task

The results of our annotation exercise are indicated in the final columns of Table 2. This column indicates the percentage of the total number of content words (NVA tagged words) for which an annotator marked at least one of the supplied senses as correct and the percentage of the total number of words assigned a synset by GermNet for which at least one of the synsets assigned was marked by an annotator as being appropriate. This is a raw measure of how well GermaNet could be used to sense tag our corpora. That is, in about 90% of the cases, if a word appears in GermaNet, then the annotators found that GermaNet provided an appropriate sense for the word as used in the corpus. While not disappointing, the numbers may seem low. In fact they are misleadingly low, as a significant proportion of these errors are not due to GermaNet at all. In section 4 we will discuss these error factors extensively.

3.4. Inter annotator agreement

An important question, however, was the degree to which the judgement of our annotators varied. We made provision for evaluating inter annotator agreement by having all the annotators tag one small subset of the short novel corpus. This subcorpus contains 431 tokens and was annotated by all five annotators. Only 170 of these tokens were assigned a list of synsets by the GermaNet. So there were 170 points the annotators could disagree on. To evaluate inter annotator agreement, we looked at whether for each of these 170 tokens any synset was marked or not by the annotators. The number of tokens that were not marked as having any acceptable GermaNet assigned synset is shown in Table 3. All five numbers are in the 95% intervall [35, 57] of the binomial distribution with n = 170 and p = 46.0 / 170 = 0.271.

Annotator	1	2	3	4	5
Token with no synset marked	40	44	56	50	40
Mean	46.0				
Variance	38.4				
Standard deviation	6.2				

Table 3: Basic statistics of annatation

It is not, of course, correct to infer from this that the annotators agree on which tokens to mark. To evaluate the more narrow question of whether our annotators agree on this we compared our annotators pairwise. Table 4 shows how many tokens can be counted in the union and intersection of two annotators' annotation records filtered for tokens that have no marked synset and in which each token was prefixed with a unique token ID. The size of the intersection gives the number of tokens that they agree on and the difference to the size of the union gives the number of tokens they disagree on. If, for example, annotator 1 and 2 completely agreed, the number of tokens would be max(40,44) = 44 in the union and min(40,44) = 40 in the intersection. If they disagreed as often as possible, the numbers would be 40+44 = 84 and 0. Table 4 gives these numbers, with the possible ranges in square brackets. The numbers seem to show quite good agreement.

A way of measuring inter annotator agreement is provided by Cohen's (1960) kappa statistic. This measure indicates the degree to which the observed agreement rate differs from chance, and is given by:

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

where P_a is the observed agreement rate and P_e is the expected chance agreement. Numbers above 0.80 are generally considered to give evidence for a good agreement, whereas numbers below 0.67 indicate poor agreement (Carletta 1996). Our κ values – indicated in the final column of Table 4 – are between or even above these standard values, indicating acceptable agreement.

We did not analyze agreement of polysemy judgements, that is, agreement on what sense should be assigned to which word (c.f. Veronis 1998), because they are irrelevant to our study. Furthermore, token counts per type are too small to get significant results. It is important to keep in mind that we were primarily interested in whether GermaNet is a rich enough lexical resource, not with whether the annotators agreed exactly on how to use it.

Annotator pair	Token in Union	 Intersection	κ
(1, 2)	48 [44, 84]	36 [0, 40]	0.81
(1, 3)	58 [56, 96]	38 [0, 40]	0.71
(1, 4)	51 [50, 90]	39 [0, 40]	0.82
(1, 5)	45 [40, 80]	35 [0, 40]	0.84
(2, 3)	61 [56, 100]	39 [0, 44]	0.69
(2, 4)	54 [50, 94]	40 [0, 44]	0.79
(2, 5)	50 [44, 84]	34 [0, 40]	0.75
(3, 4)	62 [56, 106]	44 [0, 50]	0.75
(3, 5)	58 [56, 96]	38 [0, 40]	0.71
(4, 5)	55 [50, 90]	35 [0, 40]	0.70
all five	64 [56, 107]	30 [0, 40]	0.75

Table 4: Inter annotator agreement

4. Error analysis

In order to analyze the quality and extent of GermaNet's coverage, then, we chose to further examine those tokens for which GermaNet should provide a synset, but for which no sysnset was marked by our annotators. These are the cases in which GermaNet fails to do its job. Our goal was to quantify this failure and to assess its most likely causes.

We take it to be the case that in the ideal case GermaNet would associate an appropriate sense for all occurrences of nouns, verbs, and adjectives. Given a perfect POS tagger a perfect lemmatizer, a perfect GermaNet and a perfect human annotator, every NVA-tagged word in our corpus should be marked by the annotators with at least one synset. (Perhaps *exactly* one would be more ideal; in our study we ignored this however. We were concerned that GermaNet be rich enough, not that it be too rich.)

In practice, of course, the results are not perfect. In the following we will discuss the degree to which our results deviated from the ideal. As we saw in Table 2, the number of content words which could be assigned a synset at all by GermaNet ranges from just over 83% to just under 94%. In only about 90% of the cases was one of the synsets assigned to a word by GermaNet marked as being the correct one by our annotators.

In fact, however, a large proportion of this error was introduced not by GermaNet, but by TreeTagger, which we used to lemmatize and tag our texts for part of speech. While errors in POS tagging could lead to suboptimal performance, POS tagging errors were fairly rare in our texts (as Schmid (1994) shows the tagger employed can reach an accuracy of about 97.5%). Furthermore, the kinds of errors that would be problematic in our task (mistagging of prepositions, adverbs or articles as nouns, verbs or adjectives) are the least common type. So POS tagging did not contribute significantly to the errors. Lemmatization errors, however, contribute significantly to the error rate, since every incorrectly lemmatized word resulted directly in an error: When a word is not properly lemmatized it is impossible for the human annotator to choose the correct synset, since this synset is not an available choice, as we have looked up the wrong word in GermaNet.

In order to evaluate GermaNet, then, we needed to classify our errors, so as to determine which errors were the result of GermaNet design or coverage problems and which, like lemmatization errors, were epiphenomenal.

4.1. The error classes

For purposes of our evaluation we took any NVA tagged token in our corpus to which no GermaNet synset was assigned to be an error, and we assigned each error occurrence to one of the following error classes: **lemma, particle**, **collocation, compound**, **derivation**, **auxiliary**, and **other**. The classification of errors was carried out by a single annotator (JS) using a Java-implemented GUI-tool. Each error was assigned to exactly one of the error classes. These classes were chosen because either they were a type of error that was particularly common, or because they were a type of error that the GermaNet developers had suggested might cause problems.

The error classes are described as follows:

Lemma. As mentioned, when a word is not properly lemmatized it is impossible for the human annotator to choose the correct synset, since it is not available for choice. An example of this kind of error is when the particle *mal* in "Mal wieder hat er es getan" is lemmatized as *malen*, the verb 'to draw'.

Particle. German seperable verbs, such as *vorschlagen*, contain prefixes which significantly alter the meaning of a verb (*schlagen* – "hit"; *vorschlagen* – "propose"). These verbs should be lemmatized as a single lexeme. Unfortunately in many contexts the prefix is not concatenated with the verb, as in:

Er *schlug* einen Kompromiss *vor*.
"He proposed a compromise."

This presents difficulties for lemmatizers. Very often the lemmatizer does not link the particle verb's root and prefix leading to a wrong lemmatized form, omitting the prefix (e.g. *schlagen* instead of *vorschlagen*).

Auxiliary. The verbs *sein* and *haben* (as well as certain modals and others) are also problematic. These verbs can be used simply as syntactic operators – auxiliaries – on the one hand, or as main verb on the other. As auxiliaries, there is a sense in which they should not be sense tagged (since they are not "open class"). In this group we mark those cases in which such a verb is not tagged but is recognized as being used as an auxiliary.

Strictly, speaking both **particle** and **auxiliary** errors can be thought of as lemmatization errors of a very specific type, and cannot really be attributed to GermaNet. In contrast to these we distinguished three types of errors that can be attributed to word-formation processes:

Collocation. Many words are used in a very specific sense in combination with other words (*ins Wasser fallen* to mean "cancelled", for example). In those cases in which the word to be tagged was recognized as forming part of a collocation, it was assigned to this class. While it is arguably not the task of a lexicon to account for collocations and idioms, we were interested in assessing the degree to which these are problematic.

Compound. Compounding – the formation of a new word from two or more existing words (for example

Errors class	Corpus			
	Fiction	News 1	News 2	News 3
Lemma	12.3	5.3	13.3	10.8
Particle	5.9	0	4.4	4.6
Auxiliary	25.3	21.1	22.2	21.5
Compound	11.5	31.1	11.1	32.3
Derivation	5.2	10.5	4.4	6.1
Collocation	2.2	5.3	2.2	1.5
Other	36.8	26.3	42.2	23.1
Total errors	269	19	46	62

Table 5: Distribution of errors by class and corpus (in percent)

Montagsauto) is a productive word formation process in German (as in English). As the sense to be associated with the compound is a fairly arbitrary function of the meaning of the constituent words (cf. Fanselow 1981), it is in principle difficult to provide appropriate synsets for words formed this way.

Derivation. The generation of nouns from verbs (for example *Vorbereitung* from *vorbereiten*) and the generation of diminutive forms (for example *Hündchen* from *Hund*) are productive process in German. These are somewhat more regular and might be accounted for by a GermaNet with sophisticated morphological processing (like that suggested by Kunze (1999) for particle verbs).

Finally there are the errors that fit into none of these classes:

Other. All other forms of derivation are covered by the "other coverage" default error class. The major component of this class is simply the set of words which are simply missing form GermaNet, i.e, those that should be and could be listed, but are not.

4.2. Results of error analysis

In Table 5 we present the distribution of the different type of errors by error class in each of our small corpora. It is clear there was significant variation across the corpora as to which error classes were predominant. The variation was particularly evident in the case of **lemma** and **compound** errors. The most significant class of errors was the **auxiliary** class. These were fairly uniform, accounting for between a quarter and a fifth of all errors in each of the corpora. The surprising fact that we noted in section 3, that the newspaper corpora appear to be better handled by GermaNet than the fiction corpus, gets a simple explanation: lemmatization-related errors were more pronounced in the newspaper corpus. In fact, looking only at non lemmatization-related errors, we see that the childrens fiction is, as we might expect, less error prone than the newspaper articles.

Error class	Part of Speech		
	Verb	Noun	Adjective
Lemma	3.5	23.7	13.9
Particle	13.2	0	1.3
Auxiliary	58.8	0	1.3
Compound	0	31.6	8.9
Derivation	1.8	17.1	1.3
Collocation	3.5	2.6	0
Other	19.3	25	73.4
Total errors	114	76	79

Table 6: Distributions of errors in Fiction corpus by class and part of speech (in percent)

The newspaper corpora evidenced significantly more errors that were due to the use of productive morphology. The **compound** errors were the most prominent, particularly in the newspaper corpora, although was significant variation here as well. Other **derivation** errors, however, had a relatively small share. **Collocations** though they appear in most corpora, also play a minor role.

In Table 6 the distribution of errors by POS is displayed. It is obvious why **particle** and **auxiliary** errors would be limited to verbs, as they are verb-specific error types. More interesting is the fact that errors that could be attributed to productive morphology were essentially limited to nouns and adjectives. Essentially only nouns were involved in **derivation** errors, while for adjectives (other than **lemma** errors) essentially only **compound** errors were present

5. Conclusion

Our results were very encouraging. On average 92% of the words which were tagged as verbs, nouns or adjectives were provided with at least one sense by GermaNet, and more than 83% were provided with at least one sense that was judged as the correct sense by our annotators. One of the major sources of error was, in fact, external to GermaNet: On average 15% of the content words were incorrectly lemmatized, leading to incorrect lookup. Additionally we found that many of the potential sources of coverage failure suggested by Hamp & Feldweg (1997) were indeed evident: productive morphological processes such as derivation and compounding as well as collocative uses of words accounted for a nearly 25% of the errors we noted. Particle verbs also presented problems for our annotators, as in some cases the verb was not lemmatized with its separable prefix. Clearly a more sophisticated lemmatizer could have eliminated some of these errors. In other cases productive combinations with main verbs gave rise to forms which were not covered by GermaNet. For nouns a predominant source of errors was the existence of a large number of nouns that were clearly derived via productive rules of derivation from verbs. These could, presumably, be looked up on the verbal hierarchy. Words formed via compounds were also a significant source of noun and adjective errors. Words that could not be properly tagged because they were used as part of a collocation accounted for only minority of the errors overall, however.

We also found that the effectiveness of GermaNet as used for the word-sense disambiguation task as well as the kinds of errors that were found was highly dependent on the variety of text to be disambiguated. This suggests that it is crucial that in WordNet evaluation both domain and text type be standardized, and that a variety of types be used.

Finally, many of the types of errors that we found were clearly German-language specific. This finding suggests that language-specific issues are quite important when evaluating the effectiveness of a particular WordNet and that simple cross-WordNet evaluation will likely lead to a incorrect evaluation of the value or coverage of a particular WordNet. With respect to GermaNet, our results suggest that sense-tagging using GermaNet, while quite good as it is, could be significantly improved by integrating additional morphological processing into the tagger. In particular, methods for dealing with compound words and derived words could lead to significant improvements.

6. References

Bloksma, L., P. Díez-Orzas, and P. Vossen, 1996. User requirements and functional specification of the EuroWordNet project. EuroWordNet (LE-4003), Deliverable D001, University of Amsterdam.

Carletta, Jean, 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249-254.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Dutoit, Dominique, Laurent Catherin, and Andreas Wagner, 1998. Specification of German and French Wordnets. EuroWordNet (LE4-8328), Deliverable 2D002.

Fanselow, Gisbert, 1981. *Zur Syntax und Semantik der Nominalkomposition – Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung des Deutschen*. Tübingen: Niemeyer.

Hamp, Birgit and Helmut Feldweg, 1997. GermaNet - a lexical-semantic Net for German. In: P. Vossen et al. (eds.), *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, pp. 9-15.

Kunze, Claudia and Andreas Wagner, 1999. Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database. *Sprache und Datenverarbeitung*.

Kunze, Claudia (ed.), 1999. Final wordnets for German, French, Estonian, and Czech. EuroWordNet (LE-8328), Deliverable 2D014.

Kunze, Claudia and Andreas Wagner, 1999. The German Wordnet. EuroWordNet (LE-8328), Deliverable 2D014.

Kunze, Claudia, 1999. Semantics of Verbs within GermaNet and EuroWordNet. In: V. Kordoni (ed.), *Proceedings of the ESSLLI-99 Workshop on 'Lexical Semantics and Linking in Constraint-Based Theories'*, pp. 189-200.

Landes, Shari, Claudia Leacock, and Randee I. Tengi, 1998. Building Semantic Concordances. In: Christiane Fellbaum, (ed.), *WordNet: an electronic lexical database*. MIT. Chapter 8, pp. 199-216.

Miller, G., M. Chodorow, S. Landes, C. Leacock, and R. Thomas, 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, San Francisco, Morgan Kaufmann, pp. 240-243.

Miller, George A. (ed.), 1990. WordNet: An on-line lexical database. Special issue of *International Journal of Lexicography*, 3 (4).

Resnik, Philip, 1998. *WordNet and Class-Based Probabilities.* In: Christiane Fellbaum (ed.), *WordNet: an electronic lexical database.* Cambridge: MIT Press, p. 239-263.

Schmid, Helmut, 1994, Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, September 1994.

Véronis, J., 1998. A study of polysemy judgements and inter-annotator agreement. Programme and advanced papers of the Senseval workshop, 2-4 September 1998. Herstmonceux Castle, England.

Vossen, Piek (ed.), 1999. WordNet1.5 in EuroWordNet format. Deliverable D032D033/2D014 Part B1.

Glosses in WordNet 1.5 and Their Standardization/Consistency

(The Exercise for Balkanet)

Karel Pala and Pavel Smrz

Faculty of Informatics, Masaryk University
Botanicka 68a, 60200 Brno, Czech Republic
{pala,smrz}@fi.muni.cz

Abstract

In this contribution we present an analysis of selected WN 1.5 glosses and dictionary definitions from other resources -- we examined what is the internal (semantic) organization of the glosses and dictionary definitions, i.e. how reliably and systematically they follow the standard principles of building dictionary definitions. The results following from the presented analysis should be applied in building glosses within Czech WordNet and hopefully they can serve as an exercise for other partners within Balkanet Project

1. Introduction

In this contribution we present an analysis of selected WN 1.5 glosses and dictionary definitions from other resources -- we examined what is the internal (semantic) organization of the glosses and dictionary definitions, i.e. how reliably and systematically they follow the standard principles of building dictionary definitions. The results following from the presented analysis should be applied in building glosses within Czech WordNet and hopefully they can serve as an exercise for other partners within Balkanet Project (2001).

When working on EuroWordNet (Vossen,1999) and now on Balkanet (2001) one has to have a look at the glosses in WN 1.5 and examine them quite closely since so far they have been regularly used as the references to the individual synsets or, in other words, as the descriptions of their senses. It is no secret that there are many reservations with regard to the glosses, especially to their properties from the lexicographical point of view. The main objections that can be heard are that the glosses are not consistent enough, that quite often they are nothing more than just examples and sometimes they are completely missing .

In writing dictionary definitions the following techniques are regarded as the standard ones:

1. definitions using *genus proximum* and the *distinguishers* (typically for nouns),
2. definitions using semantic components or features (verbs, e. g. **hurt***:*6 = cause pain),
3. definitions based on the relation of *troponymy* (*laugh – guffaw*),
4. definitions using synonymical explanations (typical for adjectives, e. g. *clever – smart*),
5. definitions based on collocational determination of the sense (e. g. *a bad student*, *bad debt*).
6. definitions based on the descriptions of events or situations (see e.g. definition of **bend**:1 (in Cobuild95, p.144) – *when you* **bend** *you move top part of your body downwards and forwards*,
7. definitions exploiting various kinds of *ad hoc* descriptions or explanations or just mere examples as e.g .**bring**:2 in WordNet 1.5 – *bring or fetch; "Could you bring over the wine?"*

2. Noun Glosses in WN 1.5

A large group of nouns denote the particular physical objects such as *table:1*, *chair:2*, etc. Thus we selected few "furniture" expressions and examined their glosses. It can be said that they "behave" in a relatively standard way, typically, these glosses follow the classical dictionary definition pattern, i.e. first part of the gloss consists of *genus proximum* (GP) and the second one represents the *distinguishers* (d1, d2, …, dn). Some slight inconsistencies can be observed: while **table***:*2 has as its genus proximum *a piece of furniture* (and other hyponyms of it as well), **chair***:*2 has as its GP *a seat for one person* (**seat**:2) and only then **seat**:2 displays as its GP *a piece of furniture.* Thus a question may be asked why the principle of GP is not followed strictly here. The good news perhaps is that the GP expressions in the whole WordNet can be checked and in our view corrected semi-automatically using the corresponding H/H trees. The same can hardly be applied to the distinguishers but we suggest to formalize this part of the gloss giving it a rigid structure in the form GP + *d1, d2, ..., dn.* More examples from WN 1.5 can be given, e.g.: **knife**:1 *cutting instrument + d1, d2,...* but in the corresponding H/H tree we find as the next upper node **edge tool**:1 – *any tool with sharp cutting edge.* The conclusion is obvious: we should try to avoid these inconsistencies in building Czech glosses and it can be seen that they can be checked semi-automatically as well. We examined also some other nouns like **digital computer**:1 or **house**:1 and it can be concluded that the situation with respect to their glosses is more or less the same.

The next point we are interested in is the semantic organization of the noun glosses or dictionary definitions in general, and how it is related to their syntactic structures. We can observe here quite a good parallelism between GP and the first noun group in the dictionary definition.

If we have look at the distinguishers it can be seen that they are expressed in several ways: as noun groups, relative sentences, adjectival phrases with complements or as prepositional groups. The closer examination, however, shows that the picture is more complicated and the

no of entries ("sentences" processed)	5935	100 %	
not applied	1207	20.3 %	
applied	4728	79.7 %	
from this:			
def1:	548	11.6 %	entry = one_word_synonym
def2:	2987	63.2 %	entry = (Ng | Pg)+
def3:	877	18.5 %	entry = (Ng | Pg)+ Ap (Ng | Pg)+
def4:	92	2.0 %	entry = Ng Sr
def5:	5	0.1 %	combination of def3 and def4
def6:	201	4.2 %	[kdo|co|někdo|něco] .*
def7:	17	0.4 %	[schopnost|neschopnost] .*

Table 1. Frequencies of the different definition types

corresponding surface syntactic structures are much richer (see below).

Thus it is our opinion that we should try to parse the dictionary definitions in order to discover the inventory of the syntactic structures that may correspond to the GP + d1, d2, ... dn scheme. For this purpose we again selected several typical "furniture"examples from SSJČ (1960) together with their English equivalents from NODE (1998). Angle brackets in Czech descriptions mark out the particular groups (and the grammatical cases in which they may occur).

stůl: <*kus*>$_{ng1}$ <*nábytku*>$_{ng2}$ <*tvořený*>$_{ap}$ <(*vodorovnou*) *deskou*>$_{ng7}$ <*na nohách*>p$_{png6}$ <*nebo*>$_{conj}$ <*na podstavci*>$_{png6}$
table: ***a piece of furniture*** *with a flat top and one or more legs, providing a level surface on which objects may be placed, and which can be used for such purpose as eating, writing, working or playing games*

židle: <*přenosný kus*>$_{ng1}$ <*nábytku*>$_{ng2}$ <(*s opěradlem*)>$_{ng7}$ <*k sezení*>$_{png3}$ <*pro jednu osobu*>$_{png4}$
chair: ***a separate seat*** *for one person, typically with a back and four legs*

křeslo: <*pohodlné sedadlo*>$_{ng1}$ <*s opěradly*>$_{ng7}$
armchair: ***a large, comfortable chair*** *with side supports for a person's arm*

skříň: <*vyšší kus*>$_{ng1}$ <*nábytku*>$_{ng2}$ <*na ukládání různých předmětů*>$_{png4}$ <*nebo*>$_{conj}$ <*na věšení šatstva*>$_{png4}$
cupboard: ***a piece of furniture*** *with a door and usually shelves, used for storage*

blbec, **blb**: <*velký hlupák*>$_{ng1}$, <*pitomec*>$_{ng1}$, <*idiot*>$_{ng1}$
idiot: *a stupid* ***person***

student: <*posluchač*>$_{ng1}$ <*vysoké školy*>$_{ng2}$ <*nebo*>$_{conj}$ <*žák*>$_{ng1}$ <*střední školy*>$_{ng2}$
student: ***a person*** *who is studying at a university or other place of higher education*}

2.1. Syntactic structures of the dictionary definitions

The basic Table 1 has been obtained from the sample containing 10 000 noun dictionary definitions from SSJČ and show the main types of the syntactic patterns as they can be found within the noun dictionary definitions in SSJČ.

It can be said the definitions of the entries for whose no structure has been found usually can be intuitively classified as belonging to some of the groups 1-5. However, they may display very complicated structures (e.g. very complicated attributive noun groups), that prevent the parser (the particular rules in it) from recognizing them. There are only few entries that do not belong into any of the introduced groups/categories, for example *názor*, *že ...* (*the opinion that …*)

2.2. Czech WordNet

What also can be done is to check semi-automatically the heads of these noun groups against the corresponding nouns in Czech WordNet and to see how regularly they contain the hyperonymical expressions (such as **furniture** in our example group of selected furniture nouns) – this can be done by comparing them with the corresponding H/H trees in WordNet.

If we take the parsed syntactic structures of the processed dictionary definitions and extract their head noun groups representing (according to our parser) the GP pattern we obtain a list of expressions that are hyperonyms of the headwords in the dictionary definitions. The first part of this list is given below and it contains 30 most frequent (Czech) hyperonyms (sorted according to their frequency) from our sample of dictionary definitions. To confirm that they are hyperonyms we compared them with the corresponding expressions from Czech and English WordNet (the first number indicates the frequency in the Czech sample, then there is Czech literal with its sense number in Czech WordNet and its English equivalent with its respective sense number as well. The results of the comparison show that all the expressions extracted from the dictionary definitions are hyperonymical, thus in this way confirming our starting assumption that GP patterns can be processed and obtained from the dictionary definitions automatically. Then it is also possible to check for their consistency. The next goal is to try to recognize the distinguishers at least in a semi-automatic way though we are aware that this task is not going to be as easy as the former one.

173: kdo (who)
91: přístroj:1 (apparatus:1)
72: druh:1 (sort:2, kind:1)
63: zařízení:1 (installation:2)
56: část:1 (part:3)
52: člověk:1 (human:1, person:1)
42: souhrn:1 (aggregate:1, sum:1)
42: místo:1 (place:10)
37: nástroj:2 (instrument:2)
36: obor:2 (discipline:5)
35: nauka:1 (doctrine:1)
34: látka:2 (matter:1, substance:1)
28: přísluąník:1 (member:4)
27: skupina:1 (group:1)
26: způsob:2 (means:1, way:1)
22: jednotka:3 (unit:8)
21: něco (something)
21: činnost:1 (activity:1)
20: stav:1 (state:1)
20: součást:2 (component:1)
19: vlastnost:1 (quality:1)
18: místnost:1 (room:1)
18: hornina:1 (rock:1, stone:1
17: stroj:1 (machine:2)
16: útvar:2 (formation:5)
16: sloučenina:1 (compound:4)
16: schopnost:3 (ability:1)
16: pracovník:1 (worker:2)
14: oddělení:2 (department:1)
14: nedostatek:1 (deficiency:1)
14: názor:1 (opinion:1).

3. Verb Glosses in WN 1.5

At the first glance it can be observed that the verb glosses are less consistent and regular than noun ones. Also some glosses are missing more frequently (e.g. *write:*7). We have selected verb *to kill* and its hyponyms to see how reliable the glosses are and how they are built. If we take *kill:5 cause to die* we can immediately see that GP + d1,…, dn principle does not apply here. This is generally due to the fact that the semantic nature of verbs as the relational elements is different from the nouns and that is why they require other types of definitions.

With *kill:5* the analysis to the simpler semantic components is used (type 2 above), however the problem is that the respective semantic components are used rather spontaneously, they are not defined anywhere and they are in no way related to the Top Ontology which certainly represents a collection of the specific semantic components or features. It is very instructive to examine some of the hyponyms of *kill:5* and their glosses:

behead:1 cut the head of sb (synonymical explanation)

drown:3 kill by submerging in water (troponymy relation)

poison:5 no gloss at all

shoot:16 kill by firing a missile (troponymy relation)

stone:7 "adulterers should be stoned according to the Koran" (just the example)

strangle:1 squeeze the throat of sb (synonymical explanation)

sabre: in the sense of killing not found in BNC

overlay: in the sense of killing not found in BNC

The picture we can see is rather confusing: in the cases of *drown:3* and *shoot:16* the relation of troponymy is used as the defining principle in the gloss (different manners of killing), *behead:1* and *strangle1* are defined by synonymical explanations, however *strangle:1* is not defined correctly, to squeeze the throat of a person is not enough to kill him or her, thus the gloss is defective. Moreover, for *stone:1* the example is offered instead of the definition, though *to kill by stoning* certainly could have been used. To complete this certainly not consistent view we can only add that *poison:5* has no gloss assigned at all in WN 1.5 though again *kill by using poison* offers itself as an obvious solution. It may be interesting to note that *sabre*:4 given in WordNet 1.5 as a hyponym of *kill*:5 does not occur in British National Corpus at all.

3.1. The Possible Solutions for Verbs

One of the techniques that has to be considered with regard to the verbs is an appropriate semantic classification of verbs yielding the semantic classes of verbs. The information about the semantic class a verb belongs to can become a part of the gloss/definition and can make it more systematic. Though the criteria for establishing the semantic classes may be in a certain degree arbitrary on the other hand they may be compared with Genus Proximum principle that seem to work well for nouns.

Levin's (Levin, 1993) semantic classification of English verbs appears as an interesting solution – we have tried to develop a similar semantic classification of Czech verbs that can be applied here.

4. Adjective (and Adverb) Glosses in WN 1.5

The selected examples of the adjective synsets for *good* can well demonstrate the point.

good:8, dear:2 with or in a close relationship: "a good friend"

good:10 "good taste" (an example only)

good:12 resulting favorably: "it is a good thing that I wasn't there"

good:13, unspoiled:2 "the meat is still good" (an example only)

good:14 not forged: "a good dollar bill"

good:15 having desirable or positive qualities, esp. those suitable for a thing specified: "good news from the hospital", "a good joke", "a good secretary"

good:16 morally admirable

good:23, just:6 of moral excellence: "a genuinely good person"

good:18 appealing to the mind: "good music", "a serious book"

good:19 agreeable or pleasant: "good manners"

good:25, secure:12 financially sound: "a good investment"

good:26 in excellent condition: "good teeth"

good:27 well above average in performance: "a good student"

good:29, lucky:4 "it is good that nobody saw you" (an example only)

in good taste:1 no gloss, syntactically this case can be hardly classified as an adjective.

It can be seen that for adjective *good* the definitions of the type 4, 5, 6 are used. The most frequently used are the synonymical explanations (type 4 definitions) combined with the examples of typical collocations (type 6 definitions). Only *good:16* does not include a collocational example.

The presented examples also clearly demonstrate that many senses of *good* are very close to each other and it is not easy to discriminate them. It can be observed that *good:15* seems to cover/represent the main sense of *good* and that *good:18* or *:26* or *:*27 just stress some rather arbitrarily selected semantic features such as *in excellent condition* which can be certainly classified under *a positive quality.* The adduced examples convincingly show how the senses of *good* are split into the fine grained senses but at the same time the question has to be asked what can we gain by splitting senses in this way (quite typical for WN 1.5)? The hope is that the split senses can be integrated into the larger groups and in this way the number of senses can be reasonably reduced to obtain simpler and better applicable collection of the senses. In our view the appropriate sets of the semantic features have to be considered in combination with the collocational examples – in this way the operational classification procedures (relying on corpora) for reasonably large group of adjectives and adverbs can be obtained.

The obvious conclusion also is that it is necessary to pay the more detailed attention to the collocational examples (type 6 definitions, if they can be taken as such), to explore their behaviour in the corpora and on this ground to design the techniques of their semiautomatic handling.

5. The Conclusions for Standardization

The above analysis leads us to the following steps in the building glosses within Czech WordNet (with the hope that they can appear useful in the development of other WordNets as well):

- to use the different types of definitions for the different parts of speech in a systematic way, i.e.. GP + d1, d2,…, dn mostly for nouns, semantic components and troponymy relations for verbs and synonymical explanations combined with collocational examples for adjectives,
- to use the semantic classification of Czech verbs and integrate it appropriately into the glosses,
- to examine in a more detailed way the GP + d1, d2,…, dn definitions for nouns and to check whether the distinguishers can be inherited systematically within H/H trees,
- to examine whether the distinguishers can also capture the relation of meronymy/holonymy and in the positive case to find out how frequent it is,
- to explore systematically the collocational examples using corpus data and integrate them systematically into the adjective glosses,
- the ultimate goal of the mentioned steps is to obtain the glosses for the particular synsets that would be as systematic, formal and consistent as possible.

We have tried to show how the indicated solutions may work for the selected collections of Czech synsets and in this way they may help to standardize the glosses used in Czech WordNet..

6. Bibliography

Balkanet Project, 2001, www pages http://www.ceid.upatras.gr/Balkanet/

Collins Cobuild English Dictionary, ed. by J. Sinclair, London, Harper Collins Publishers, 1995.

Havránek B. et al., Slovník spisovného jazyka českého (SSJČ, Dictionary of Written Czech), Academia, Praha, 1960.

Levin, B., English Verb Classes and Alternations, The University of Chicago Press, Chicago, 1993.

New Oxford Dictionary of English, ed. by P. Hanks, Oxford University Pres, Oxford, 1998.

Vossen, P., EuroWordNet 1, 2, Final Report, University of Amsterdam, CD ROM, 1999.

Žáčková, E.: Partial Parsing (of Czech), Ph.D. Thesis, Masaryk University, Brno, 2002.

Standardizing Wordnets in a Web-compliant Format: The Case of GermaNet

Claudia Kunze, Lothar Lemnitzer

Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 113, 72074 Tübingen, Germany
{kunze,lothar}@sfs.uni-tuebingen.de

Abstract

Following the success of the Princeton WordNet, a range of wordnet initiatives have been launched, either monolingual or multilingual. The variety of wordnets which have a common core architecture but also their language-specific peculiarities calls for a common standard to enhance interoperability, to merge of different lexical resources and to define a common application programme interfaces. At the same time, the drive for the "semantic web" and the resp. need for ontologies calls for XML- and RDF-binding of at least the common core architecture of wordnets. The GermaNet group therefore wishes to contribute to the standardization of wordnet architectures by presenting the data model of GermaNet, an XML binding of this data model and some proposals for a common terminology.

1. Introduction

Have you ever tried to use your razor or your hair dryer in another country than that where you bought this device? Even in Europe you might be caught in a situation where the plug of your device and the socket in your hotel room are incompatible. You might end up buying an expensive adapter at the reception desk of your hotel. Missing standards can be a burden or even an obstacle to further development.

Avoiding a waste of time and money is one incentive of undergoing the effort of negotiating a standard, which in itself can be a time-consuming task.

The situation of the European traveller might be comparable to that of a language engineer who wants to:

- Use wordnets of various languages in a multilingual application environment
- Adapt an application which uses a wordnet in one language to another language and wants to adapt an available wordnet for that language, too
- Couple a dictionary management or visualization tool for a wordnet in one language with the wordnet of his / her language (see Pavelek and Pala, this volume)

It is therefore in our opinion worth the effort to discuss the following issues. In what manner are the WordNet architecture, the EuroWordNet architecture and the architecture of any individual wordnet are related? Is there a common core architecture? Do we really mean the same if we use the same concepts and terms to describe our resources? Do we perhaps refer to the same concepts though we are using different terms?

The GermaNet development group wants to contribute to this discussion. First of all, we describe the features which GermaNets shares with other wordnets, in particular the Princeton WordNet (section 2). We will present the data model of GermaNet in an application neutral graphical form, using the Entity Relationship model (section 3), as well s an XML binding of the GermaNet data model (section 4). In section 5 we will show a way of integrating the Interlingual Index of the EuroWordNet architecture into the GermaNet architecture. We will explicate the terminology we use and relate it to other wordnet terminologies, the Princeton WordNet and the Czech word net in particular (section 6). Finally, we will raise compatibility issues and suggest solutions to at least some of them (section 7).

The task we are facing is not exciting nor is it easy. Anyway, our motivation to solve it should be clear to all developers of wordnets: Think of the plug and the socket!

2. GermaNet: its standard core and its peculiarities

2.1. General Remarks

The fundamental lack of electronic lexical-semantic resources for German (see Hamp & Feldweg (1997)) was the major motivation for constructing GermaNet a few years ago. Therefore, a first project (SLD) created an online thesaurus covering the German basic vocabulary. GermaNet adopted the design principles and the database technology from the Princeton WordNet. However, GermaNet includes principle-based modifications on the constructional and content-oriented level which we will describe later on.

GermaNet currently covers some 40,000 synsets with more than 60,000 word meanings, modelling nouns, verbs, adjectives and adverbs (see Kunze (2001)). Within the EuroWordNet project, GermaNet was integrated into the polylingual EuroWordNet database (see Vossen (1999), Wagner and Kunze (1999)). We followed the merge approach, i.e., a wordnet is built independently from WordNet and the synsets are linked to the Interlingual Index (ILI) by creating the appropriate relations. The merge approach preserves language-specific patterns with differing hierarchical structures in comparison to the WordNet structure.

2.2. Major differences to WordNet

In spite of its general similarity with and compatibility to WordNet, we can state the following differences for GermaNet:

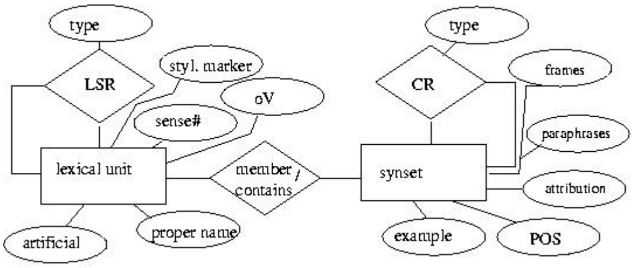
- we are using **artificial**, i.e. non-lexicalised **concepts**, which have been introduced to fill

- lexical gaps, to balance the taxonomical structure more adequately and to avoid unjustified co-hyponymy;
- in GermaNet, **adjectives** are ordered hierarchically as opposed to Princeton's grouping by the satellite approach;
- we pursued a uniform treatment of **meronymy** within GermaNet, whereas WordNet has established three different pointers for *Part, Member* and *Substance*;
- within GermaNet, the **causation relation** can be encoded between all parts of speech, not only between verbs and adjectives;
- due to emphasizing the syntax-semantics-interface for disambiguation tasks we accounted for over one hundred verbal **subcategorisation frames**. These frames are more elaborate than the WordNet frames, and, furthermore, for each verb reading we provide a typical **example**.

These differences and their technical impact on compatibility for the XML conversion are outlined in more detail below.

3. The data model of GermaNet

We visualize the data structure by graphic means using the Entity-Relationship Model (Chen, 1976).



CR=conceptual relation; LSR=lexical–semantic relation;oV=orthographic variant

Fig. 1: Entity-Relationship graph of the GermaNet data model

The graph in figure 1depicts:

- the **objects**, synsets and lexical units, which are represented as *rectangles*,
- the **attributes** of these objects, represented as *circles*,
- the **relations**, represented as *diamonds*. In GermaNet, like in WordNet, we distinguish:
    - o **conceptual** relations (CR) which hold between instances of the synset object (e.g. hyperonymy) from
    - o **lexical-semantic** relations (LSR) which hold between instances of the lexical unit object (e.g. antonymy).

From an Entity-Relationship model, one can formally derive the conceptual structure of a relational database in a normalized form (Seesing, 1993). One can also, however not as unambiguously, derive a DTD or schema for an encoding of the data which is in line with the XML standard.

4. An XML Binding of the Data Model

We have converted the GermaNet data into a set of XML-encoded documents which conform to two *Document type definitions* (DTDs). One DTD represents the objects (synsets and lexical units) and their attributes, the other represents the relations between these objects.

In the following, we will describe both DTDs. The first DTD represents the data model of the objects and their attributes. It is recorded completely in fig. 2.

```
<!-- DTD for Germanet objects -->
<!-- Version 1.9, March 2002 -->>
<!-- Copyright: Sem. f. Sprachwissenschaft der
Universität Tübingen -->

<!ELEMENT synsets      (synset)+>
<!ELEMENT synset      ((lexUnit)+, attribution?,
frames?, paraphrases?, examples?)>
<!ATTLIST synset       id             ID     #REQUIRED
                    wordClass        CDATA   #IMPLIED
                    lexGroup         CDATA #IMPLIED>
<!ELEMENT lexUnit      (orthForm)+>
<!ATTLIST lexUnit      id            ID     #REQUIRED
               StilMarkierung  (ja|nein)      "nein"
               sense  CDATA           #REQUIRED
               orthVar        (ja|nein) "nein"
               artificial     (ja|nein)  #REQUIRED
               Eigenname      (ja|nein) #REQUIRED >
<!ELEMENT orthForm      (#PCDATA)>
<!ELEMENT paraphrases   (paraphrase)+>
<!ELEMENT paraphrase    (#PCDATA)>
<!ELEMENT examples      (example)+>
<!ELEMENT example       (text, frame*)>
<!ELEMENT frames        (frame)+>
<!ELEMENT attribution   (#PCDATA)>
<!ELEMENT text          (#PCDATA)>
<!ELEMENT frame         (#PCDATA)>
```

Fig 2: The GermaNet objects DTD

Description: Documents which conform to this DTD contain a set of *synsets*. Every *synset* consists of at least one *lexical unit*. *Paraphrases* may be given to characterize the meaning of the synset and an *attribution* as well as *examples* may be added to illustrate the use of its member lexical units. For verb synsets, subcategorization *frames* are given. The individual lexical units are characterized by a set of attributes, e.g. *sense* number and stylistic marker (*StilMarkierung*). A concept can be represented by a string which does not correspond to a lexical unit in the German vocabulary. Such a unit will be marked as *artificial*. The content model of most atomic elements is set to *#PCDATA*, therefore minimizing data type restrictions. It is up to the lexicographers to fill the elements with appropriate data.

```
<!-- DTD for GermaNet relation files.-->
<!-- Version 1.4, März 2002 -->>
<!-- Copyright: Sem. f. Sprachwissenschaft der
Universität Tübingen -->

<!ELEMENT relations (lex_rel | con_rel)+>
```

```
<!ELEMENT lex_rel (locator+, arc+)>
<!ATTLIST lex_rel name (antonymy | pertonymy |
participleOf) #REQUIRED
              dir (one | both) #REQUIRED
              sense          CDATA #REQUIRED
              xmlns:xlink CDATA #FIXED
'http://www.w3.org/1999/xlink'
               xlink:type (extended) #FIXED 'extended'>
<!ELEMENT con_rel (locator+, arc+)>
<!ATTLIST con_rel name (hyperonymy | meronymy |
holonymy | entailment | causation | association)
#REQUIRED
              dir (one | both) #REQUIRED
              xmlns:xlink CDATA #FIXED
'http://www.w3.org/1999/xlink'
              xlink:type (extended) #FIXED 'extended'>
<!ELEMENT locator EMPTY>
<!ATTLIST locator xlink:type (locator) #FIXED 'locator'
                xlink:href CDATA #REQUIRED
                xlink:label CDATA #REQUIRED>

<!ELEMENT arc EMPTY>
<!ATTLIST arc xlink:type (arc) #FIXED 'arc'
          xlink:from CDATA #REQUIRED
          xlink:to CDATA #REQUIRED
          xlink:actuate (onRequest) #FIXED 'onRequest'
          xlink:show (other) #FIXED 'other'>
```

Fig. 3: The GermaNet relations DTD

Description: Documents which conform to this DTD contain a set of relations which are either conceptual or lexical relations. These relations are characterized by their type (attribute: name) and they are marked as either symmetrical or directed (attribute: dir). They are realized as links according to the XLink specification: a link consists of two nodes (locators, specified through the IDs of the synsets or lexical units) and one or two arcs, depending on whether the relation is directed or symmetrical. The attributes of the 'arc' element specifies the processual behaviour whenever a link is traversed.

5. Extensions of the Data Model and DTD

5.1. Cross-lingual extension with EuroWordNet

Within a European project, the wordnets of several languages, including German, have been integrated into the polylingual architecture of the EuroWordNet database. This has been achieved by linking the language-specific concepts to the Interlingual Index (ILI) of EuroWordNet (Vossen, 1999). The ILI has the following features:

- It is an unordered list of synsets, so-called ILI-records;
- Each ILI-record has a unique identifier, consisting of a categorial marker and a sense ID;
- The ILI-records have basically been derived from the Princeton WordNet; some new ones have evolved from the project;
- The ILI does not account for structural relations between the records. The structural relations are provided by the language-specific wordnets being linked to the ILI.

An example of the ILI and its satellites is shown in fig. 4

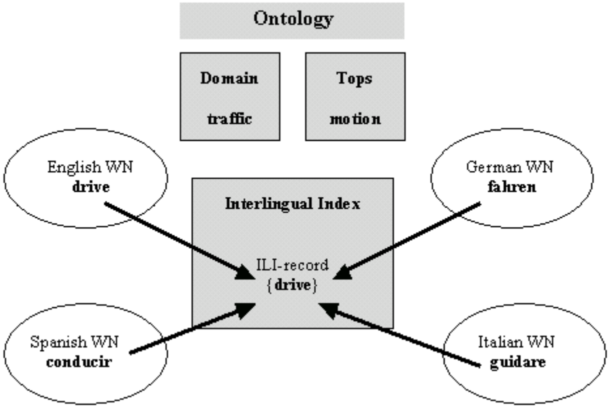


Fig 4: Partial architecture of the EuroWordNet database

From fig. 6, one can derive that there is no direct connection between the wordnets of the various languages. Mappings between language-specific wordnets are mediated by the Interlingual Index.

The following inventory of equivalence relations for connecting synsets of an individual wordnet to the ILI is provided by the EWN specification:

- EQ_SYNONYM
- EQ_NEAR_SYNONYM
- EQ_HAS_HYPERONYM
- EQ_HAS_HYPONYM
- EQ_INVOLVED
- EQ_ROLE
- EQ_IS_CAUSED_BY
- EQ_CAUSES
- EQ_HAS_HOLONYM
- EQ_HAS_MERONYM
- EQ_HAS_SUBEVENT
- EQ_IS_SUBEVENT OF
- EQ_BE_IN STATE
- EQ_IS_STATE_OF

Furthermore, the relations between a wordnet synset and an ILI element are directed. The wordnet synset is the source and the ILI element is the target of this link.

Given these characteristics, we extend the GermaNet relations DTD in the following way:

- Introduce an additional element for this new class of links ("equivalence link")
- Characterize the link as directed
- Define an attribute with the closed set of types which characterize ILI links in the EuroWordNet architecture
- Define two locators for the link, one of which must have an identifier designating a GermaNet synset, the other an identifier designating an ILI element
- Define an arc between these two locators and specify the application semantics of the link during traversal of this arc.

The result of this procedure is shown in fig. 5.

```
<!-- DTD for GermaNet relation files – extended,
interlingual version.-->
....
<!ELEMENT relations (lex_rel | con_rel | eq_rel)+>
…
<!ELEMENT eq_rel (locator+, arc+)>
<!ATTLIST eq_rel name (EQ_SYNONYM|
EQ_NEAR_SYNONYM| EQ_HAS_HYPERONYM|
EQ_HAS_HYPONYM| EQ_INVOLVED| EQ_ROLE|
EQ_IS_CAUSED_BY| EQ_CAUSES|
EQ_HAS_HOLONYM| EQ_HAS_MERONYM|
EQ_HAS_SUBEVENT| EQ_IS_SUBEVENT OF|
EQ_BE_IN STATE| EQ_IS_STATE_OF) #REQUIRED
            dir (one | both) #FIXED 'one'
            xmlns:xlink CDATA #FIXED
'http://www.w3.org/1999/xlink'
            xlink:type (extended) #FIXED 'extended'>
…..
```

Fig. 5: Extended interlingual relations DTD

A core of GermaNet synsets has been linked to the Interlingual Index (ILI). In the process of linking these synsets have got a separate ID. We could have used the IDs as a key to those synsets. The fact however that only one third of the synsets is linked to the ILI led us to the decision to employ our own scheme of IDs, which are processed on conversion of the data. We will provide a mapping from ILI link IDs to the IDs generated by our programs.

6. Terminology

In this section we want to compare the terminology we use with those employed by other wordnet development groups. The documents we refer to are the description of the Princeton WordNet (Miller, Fellbaum) and the description of the Czech WordNet (Pavelek and Pala, this volume).

Uncontroversially the *synset* is the central object of every wordnet. A synset consists of one or many members. In the RDF binding of WordNet these members are called *word forms*. Some wordnet development groups call them *synonyms*. We decided to use neither because:

- *Word form* denotes a concrete linguistic entity, in many times inflected and found in texts, whereas the members of synsets are lexical abstractions which are represented by one form, the so called base form.
- *Synonym* is a genuinly relational term. A lexical sign can be a synonym only in relation to some other lexical sign.

In contrast, we use the tem *lexical unit* to establish a distinct kind of object which has its own attribute-value pairs. Furthermore, the term is also used with traditionally organized lexical resources and can therefore facilitate a merge of different kinds of lexical resources.

Lexical units are organized in synsets by the central relation of *synonymy*. It is however not clear to us wether all groups employ the same definition of *synonymy* and the same set of operational tests. On the other hand, the linking of synsets with a narrow definition of *synonymy* to synsets with a wider definition of it – in interlingual relations – might cause severe problems in multi-lingual application environments. We believe that the reliability of equivalence relations between synsets is worth testing.

Lexical units are represented by “literal strings” (we are using the term *orthographical form*) and sense numbers.

Part of speech plays a central role as a feature of synsets in that it divides the set of concepts into subsets. Most wordnets comprise nouns, verbs and adjectives. There is a strong tendency therefore to stick to these parts of speech even if they do not prove adequate for all languages (see Kahusk, this volume, for a more detailed discussion).

Most wordnets provide a textual description of synsets. In WordNet and in the Czech word net they are called *glosses*, whereas we are using the term *paraphrase.* The WordNet RDF *glossary* however seems to comprise paraphrases and examples, which are two different data types in GermaNet. This point needs clarification. We are not against using the term *gloss* if it is well defined.

Again there is little difference in the kinds and types of relations within wordnets. There are *conceptual relations* between synsets and *lexical-semantic relations* between lexical units. Some wordnet development groups however (see Pavalek and Pala, this volume; Vider, this volume) use the tem *semantic relations* instead of *conceptual relations.* The Czech wordnet developers are using *literal relations* to signify what we call *lexical-semantic relations.*

In addition, EuroWordNet 1 defined a set of interlingual relations between synsets on which at least the members in this project phase agreed. Furthermore this project provided a proposal for a set of intralingual relations which at least some of the new members of the wordnet society in Europe have taken over.

The Estonian wordnet applies a much richer set of semantic relations than e.g. WordNet(see Vider et al., 1999). Furthermore the developers are in need of a set of subtypes for the EWN relation *derived / has_derived / derived_from.*

Furthermore there are differences between the architecture of WordNet (at least the RDF binding), GermaNet and the Czech wordnet.

In WordNet (the RDF version), synsets, glosses and relations (or to be precise, the hyperonymy relation) are organized in different files. In GermaNet (the XML version) the synsets and synset related features are organized separately from the relations, which are called *links* in orientation to the Xlink standard. In the Czech wordnet synsets and relations are organized in one data structure. Glosses are stored in a different file for the simple reason that the English WordNet glosses are used until Czech glosses will be generated (see Pavelek and Pala, this volume). This however, seems to be a minor, merely technical point. At least, GermaNet offers a data structure comparable to the Czech wordnet.

There are only a few if any information types other than the the ones mentioned which are shared by a larger number of wordnets. Subcategorization frames seems to be one candidate. However it might be even more difficult to come to an agreement about the status and the information provided by this data type. This and other information should be treated as particular to any individual wordnet.

7. Compatibility issues

In this section, we will raise several compatibility issues and show how they can be solved within the XML framework We will elaborate on six types of structural differences between WordNet and GermaNet:

1. Objects or relations might have different extensions in both nets, as is the case with the CAUSE relation. In WordNet, this relation holds exclusively between verbs and adjectives. In GermaNet, synsets of all word classes are in the domain of this relation. True compatibility would require a finer granularity of the CAUSE relation in GermaNet. This could be realised by adding an attribute to it. The values of this attribute would lead to at least two subsets of items: one which is extensionally identical with the WordNet CAUSE relation and one which characterises the GermaNet-specific extension.
2. The granularity of a relation differs. For example, WordNet divides the generic part-whole relation into three sub-relations: part (e.g. *arm,body*), member (e.g. *director, staff*), substance (e.g. *glass, glass plate*). Other values might be added to this list. GermaNet, in contrast, uniformly applies the generic relation. We recommend for WordNet or any other wordnet which applies this architecture to add an attribute to a truly generic part-whole-relation which divides the instances into three classes. In GermaNet, this attribute might get a value ANY, until a more fine-grained specification is implemented.
3. There are a few attributes specific to GermaNet, e.g. *StilMarkierung* (=stylistic marker) as an attribute of lexical units. For instance, the German concept *schlafen* (=sleep) has *ratzen*s, pennen*s, knacken*s, pofen*s* as hyponyms which are stylistically marked. These attributes can be INCLUDED in GermaNet and EXCLUDED elsewhere. The same holds for language-specific features of other word nets, e.g. features like *katharevousa* and *demotiki* in Greek.
4. An attribute which is equivalent in both wordnets specifies a different set of values. This holds for the *verb frame* attribute. The German verb frames which are implemented in GermaNet are a closed class. For type checking, it could have been more elegant to define an attribute with a fixed set of values. For compatibility reasons, however, we voted for an element group "frames" with frames as its elements and #PCDATA as data type
5. The adjective domain in GermaNet differs fundamentally from that in WordNet. The domain is ordered hierarchically in GermaNet, whereas WordNet applies an associative similarity relation which groups adjectives in equivalence classes. At present, we do not see any easy solution which would preserve compatibility in this case.

8. Conclusion

We presented the GermaNet data model and an XML binding for it in order to contribute to the difficult process of establishing a standard for at least the core architecture of wordnets. On the way to a standard both conceptual and terminological issues arise. With respect to visualization tools and the semantic web we decided to choose XML in general, and two DTDs in particular, to present our view of the GermaNet architecture.

9. Acknowledgments

The work we report here is in part the outcome of some student projects. We would therefore like to thank Iris Vogel (Heidelberg) and Holger Wunsch (Tübingen) for their valuable contributions. Research on GermaNet was funded by the Land Baden-Württemberg.

10. References

Chen, P. P.-S., 1976. The Entity-Relationship Model - Towards a Unified View of Data. *ACM TODS 1* No. 1 (March 1976):9-36.

Fellbaum, C., 1998. *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Hamp, B. and Feldweg, H., 1997. GermaNet - a Lexical-Semantic Net for German. In: *Proceedings of the ACL/EACL-97 workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP applications*. Madrid, July 7-12, 1997.

Kahusk, Neeme, 2002. A Lexicographer's Tool for Word Sense Tagging According to WordNet. In: *Proc. of the LREC Workshop on Word Net Structiure and Standardization and how these Affect Wordnet Applications and Evaluation, Las Palmas, 28 May 2002.*

Kunze, C., 2001. Lexikalisch-semantische Wortnetze. In K.-U. Carstensen et al. (eds.): *Computerlinguistik und Sprachtechnologie: eine Einführung*. Heidelberg; Berlin: Spektrum, Akademischer Verlag, S. 386-393.

Kunze, C. and Wagner, A., 2001. Anwendungsperspektiven des GermaNet, eines lexikalisch-semantischen Netzes für das Deutsche. In Lemberg, I. & B. Schröder & A.Storrer (eds.): *Chancen und Perspektiven computergestützter Lexikographie*. Tübingen: Niemeyer. Lexicographica Series Maior 107. S. 229-246.

Lemnitzer, L. and Kunze, C., 2002. Adapting GermaNet for the Web. *Proceedings of the first Global WordNetConference*, Central Institute of Indian Languages. Mysore, India, 2002, pp. 174-181.

Miller, G. et al., 1990. *Five papers about on WordNet.* CSL-Report, Vol. 43. Cognitive Science Laboratory, Princeton University.

Pavelek, Tomas and Pala, Karel, 2002. WordNet Standardization from a Practical Point of View. In: *Proc. of the LREC Workshop on Word Net Structiure and Standardization and how these Affect Wordnet Applications and Evaluation, Las Palmas, 28 May 2002.*

Seesing, Paul R., 1993. *Basic Systems Analysis Tools for Computer Users* (http://www.open.org/~prslkg/ syintro.htm)

The Semantic Web Community Portal (URL http://www.semanticweb.org/)

The Semantic Web Community Portal – Library (URL http://www.semanticweb.org/library)

Vider, Kadri, Paldre L., Orav, H and Õim, H, 1999. The Estonian Wordnet. In: Kunze, C.. editor*, Final Wordnets for German, French, Estonian and Czech*. EuroWordNet (LE-8328), Deliverable 2D014.

Vossen, P., ed., 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

Vossen, P., 1999. EuroWordNet. Building a Multilingual Database with Lexical-Semantic Networks for the European Languages. *Proceedings of EUROLAN'99, 4th European Summer School on Human Language Technology*. Iasi, Romania. July 19-31, 1999.

Wagner, A. and Kunze, C., 1999. Integrating GermaNet into EuroWordNet, a Multilingual Lexical-Semantic Database. *Sprache und Datenverarbeitung, SDv* Vol. 23.2/1999:5-20.

Xlink 1.0. (URL=http://www.w3.org/TR/2001/REC-xlink-20010627/)

XML 1.0. (URL=http://www.w3.org/TR/1998/REC-xml-19980210/)

WordNet Standardization from a Practical Point of View

Tomas Pavelek, Karel Pala

Faculty of Informatics, Masaryk University
Botanicka 68a, 60200 Brno, Czech Republic
{xpavelek,pala}@fi.muni.cz

Abstract

This article deals with standardization of WordNet data in practice. The format of WordNet databases mentioned here comes as a result of connection with a tool VisDic which enables browsing and editing electronic readable dictionaries. Later in the article, a possibility of embedding WordNet in a common dictionary using this format is described. Finally, a short overview of VisDic tool is presented.

1. Motivation

When we thought about the improvements of WordNet databases we knew that it would be necessary to make a tool which easily enables editing synsets, their relations and links to other wordnets. There has been a tool which met the requirements - Polaris (Louw, 1998). But Polaris was a good WordNet editor just at a first glance. It displayed many functions, but unfortunately also several serious disadvantages: It was a closed project, it was aimed to WordNet databases only and it used its own format for representation of synsets – Import/Export format (Louw, 1998).

What is needed is a program that would enable users to search also in other databases and sources like monolingual and bilingual dictionaries, dictionaries of synonyms or corpora. Thus we needed to find a format for description of possibly any type of lexical resources. The XML format fits perfectly for these purposes. In the next section, we describe the details of this format and motives that led us to the design mentioned below.

2. Format of WordNet Data

A representation of WordNet data comes from the idea that the WordNet database is a dictionary consisting of entries which correspond to the individual meanings. The meaning is described by a set of words. A meaning can also have a gloss consisting of a free text definition of the meaning. The meaning is further derived from synset relations connecting them. There are two types of synset relations (Miller 1993, Vossen 1999):

Internal Language Relations that connect synsets in the range of one language, e.g. hyperonymy, hyponymy, meronymy, holonymy relations.

External Relations which connect synsets among more languages, e.g. EQ_SYNONYM, EQ_HYPERONYM, EQ_HYPONYM.

At this point, it is necessary to make clear how to represent the relations effectively in the computer. The most effective way is to assign a key to each synset which uniquely identifies it. Then the synset can be easily referred by others just by specifying its key. But in WordNet, there already is a value which can be understood as a key, particularly, it is the Interlingual Index (ILI). Then, all relations can be represented just by their names and ILI of the target synset. Moreover, ILI immediately defines the EQ_SYNONYM relation.

The format is further extended by an information about the part of speech of each synset. The next extension divides words in the synset to a literal part and sense number part. The reason lies in the fact, that about 22% of words (e.g. *page*) have more meanings and then it is useful to distinguish them by a sense number.

Fig 1. shows the selected parts of the just described synset representation (VisDic definition). Each row contains a specific information about a synset represented by a tag. The first column contains a level of the tag in a structure. Every tag belonging to a specific level N can be understood as a part of the nearest upper tag having a level N-1. The second column contains a name of a tag. The third column contains its minimal number of repeating in a structure and the fourth column its maximal number of repeating in a structure (–1 means infinity). The fifth column contains the following information about the type of a tag:

N – the tag contains a normal text value

K – the tag contains a key value uniquely identifying the synset, this key can be used by all L, R, and E tags whose definitions follow

L – the tag contains a link to another synset, it is representing a semantic relation

R – is similar to L, but it is not necessary to store the tag, because it can be reversibly inferred by a tag stored in the sixth column

E – the tag contains an information stored in another dictionary, a name of an external tag is contained in the sixth column and a name of a dictionary in the seventh column.

Fig. 2 shows the corresponding DTD. At the first sight you can see the difference between these two descriptions. VisDic definition does not contain any information about attributes. Therefore, all tags are understood as elements from a DTD point of view. On the other side it is not crucial to specify which information should be understood as an attribute and which as an element, because all elements which have not any children should be considered as attributes from a low level processing. VisDic definition can be thus understood as a description of XML which comes from a binary representation of a database, it exactly describes a tree structure of XML, while DTD defines more data types and is more readable for humans. The difference between VisDic definition and DTD is comparable to a difference between C and Prolog programming languages.

VisDic tool that will be described later uses the simplified VisDic definition of an XML database.

```
0 SYNSET                1       1 N
  1 ILI                 1       1 K
  1 POS                 1       1 N
  1 GLOSS               0      -1 E WORD_MEANING.GLOSS    wn/ili/wn_ili
  1 SYNONYM             1       1 N
    2 LITERAL           1      -1 N
      3 SENSE           1       1 N
  1 BE_IN_STATE         0      -1 L
  1 STATE_OF            0      -1 R SYNSET.BE_IN_STATE
  1 CAUSES              0      -1 L
  1 IS_CAUSED_BY        0      -1 R SYNSET.CAUSES
  1 HYPERONYM           0      -1 L
  1 HYPONYM             0      -1 R SYNSET.HYPERONYM
  1 HOLONYM             0      -1 L
  1 MERONYM             0      -1 R SYNSET.HOLONYM
  1 SUBEVENT            0      -1 L
  1 IS_SUBEVENT_OF      0      -1 R SYNSET.SUBEVENT
  1 ANTONYM             0      -1 L
  1 INVOLVED            0      -1 L
  1 ROLE                0      -1 R SYNSET.INVOLVED
  1 XPOS_NEAR_ANTONYM   0      -1 L
  1 XPOS_NEAR_SYNONYM   0      -1 L
  1 EQ_HOLONYM          0       1 L
  1 EQ_MERONYM          0       1 R SYNSET.EQ_HOLONYM
  1 EQ_HYPERONYM        0       1 L
  1 EQ_HYPONYM          0       1 R SYNSET.EQ_HYPERONYM
```

Fig 1. VisDic definition of synset representation (selected tags)

```
<!ELEMENT SYNSET  (POS,GLOSS,SYNONYM+)>
<!ELEMENT POS     (#PCDATA)>
<!ELEMENT GLOSS   (#PCDATA)>
<!ELEMENT SYNONYM (#PCDATA,SENSE)>
<!ELEMENT SENSE   (#PCDATA)>
<!ATTLIST SYNSET  ILI               ID #REQUIRED>
<!ATTLIST SYNSET  BE_IN_STATE       IDREFS>
<!ATTLIST SYNSET  STATE_OF          IDREFS>
<!ATTLIST SYNSET  CAUSES            IDREFS>
<!ATTLIST SYNSET  IS_CAUSED_BY      IDREFS>
<!ATTLIST SYNSET  HYPERONYM         IDREFS>
<!ATTLIST SYNSET  HYPONYM           IDREFS>
<!ATTLIST SYNSET  HOLONYM           IDREFS>
<!ATTLIST SYNSET  MERONYM           IDREFS>
<!ATTLIST SYNSET  SUBEVENT          IDREFS>
<!ATTLIST SYNSET  IS_SUBEVENT_OF    IDREFS>
<!ATTLIST SYNSET  ANTONYM           IDREFS>
<!ATTLIST SYNSET  INVOLVED          IDREFS>
<!ATTLIST SYNSET  ROLE              IDREFS>
<!ATTLIST SYNSET  XPOS_NEAR_ANTONYM IDREFS>
<!ATTLIST SYNSET  XPOS_NEAR_SYNONYM IDREFS>
<!ATTLIST SYNSET  EQ_HOLONYM        IDREF>
<!ATTLIST SYNSET  EQ_MERONYM        IDREF>
<!ATTLIST SYNSET  EQ_HYPERONYM      IDREF>
<!ATTLIST SYNSET  EQ_HYPONYM        IDREF>
```

Fig 2. DTD of WordNet Database (selected tags)

3. Advantages of VisDic definition

Looking at Fig. 3 we can see the example of two synsets stored in a typical database: {psychological feature:1} (P) and {cognition:1, knowledge:1} (C). Notice the following facts:

HYPERONYM tag of C contains exactly the same ILI value (*00012517-n*) as is present in ILI tag of P. This implies, that P is a hyperonym of C. A searching of the hyperonym is reduced just to a looking up a single value *00012517-n*, which **reduces the time for searching**.

There is no HYPONYM tag. An information that C is a hyponym of P is already present in the fact, that P is a hyperonym of C. Searching for all hyponyms of P is then converted to looking up all synsets having their HYPERONYM value the same as ILI value of P. In most cases, the synset has only one hyperonym, but it can have tens of hyponyms. Then, using reversible tags as HYPONYM **reduces the size of a database**.

If we look at Fig. 1, we can see that the gloss is present in the external file called *wn/ili/wn_ili*. There is a good reason for that. There are wordnets that do not have their own glosses at the time. Until these glosses will be added, it is good to use glosses which already exist even in another language. Therefore this external link points to a special file, where all English glosses are stored. All wordnets then can point to this place and automatically load a gloss when necessary - the format allows to **link dictionaries**.

If it is necessary the user can add its own tag, such as gloss in his language, to his own WordNet and the changes take effect immediately without a need of any further processing, such as a recompilation of the WordNet database - the format is **easily extensible**.

```
<SYNSET>
  <ILI>00012517-n</ILI>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>psychological feature
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
</SYNSET>
```

```
<SYNSET>
  <ILI>00012878-n</ILI>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>cognition
      <SENSE>1</SENSE>
    </LITERAL>
    <LITERAL>knowledge
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
  <HYPERONYM>00012517-n</HYPERONYM>
</SYNSET>
```

Fig. 3. Two synsets represented in VisDic definition

4. WordNet Embedded in Another Dictionary

There are very few relations which cannot be represented in VisDic format definition: DERIVATION, ANTONYM (for literals only), IS_DERIVED_FROM, HAS_DERIVED, PERTAINS_TO, IS_PERTAINED_TO, HAS_INSTANCE and BELONG_TO_CLASS. The reason is, that these relations do not connect two synsets, but in the most cases two literals. Synsets are uniquely identified by their ILI values, but there is no way to refer to their parts – particularly literals. Although most of WordNet data do not contain these relations it is necessary to think about how they can be represented.

One possible solution of this problem would be to specify the second key (in VisDic definition say the type of a tag K2), which makes a unique identification of each literal and sense pair. The literal relations can be then labelled by L2 or R2 type of a tag. The difference between synset links and literal links should be then distinguished (synset links have L and R type of tags). The corresponding part of VisDic definition of synonyms from Fig.1 should be then replaced by data represented by Fig.4.

```
1 SYNONYM                1       1 N
  2 LITERAL              1      -1 N
    3 ID                 1       1 K2
    3 SENSE              1       1 N
    3 IS_DERIVED_FROM    0      -1 L2
    3 HAS_DERIVED        0      -1 R2 SYNSET.SYNONYM.LITERAL.IS_DERIVED_FROM
    3 DERIVATION         0      -1 L2
    3 PERTAINS_TO        0      -1 L2
    3 IS_PERTAINED_TO    0      -1 R2 SYNSET.SYNONYM.LITERAL.PERTAINS_TO
    3 ANTONYM            0      -1 L2
```

Fig. 4. Additional information in VisDic definition for literal relations

It looks quite well but now consider, that we would like to link WordNet database to another common dictionary. Most of common dictionaries are sorted by words which correspond to the literals rather than the word meanings. Word meanings in WordNet typically contain more literals. Therefore, if we would like to refer to the word *cognition* from an example in Fig. 3 we should use a SYNSET.SYNONYM.LITERAL.ID tag instead of SYNSET.ILI tag, because ILI value comprises both *cognition* and *knowledge* literals.

Now think about the real situation, when both WordNet and the other dictionary are being edited. While ILI values of synsets are strictly given, the literals' ID's are very often modified. E.g., when a user deletes the literal from a synset, adds another literal to this synset and finally realizes that the first one was correct and replaces it back, the ID will not be the same. Therefore, during every simple change in a synset, it is necessary to update all the references to the literal. It is possible to maintain ID numbers within WordNet as a compact dictionary, but it is hard to keep consistent more different dictionaries. In our view, this is one of two reasons why this approach should not be followed.

The second reason is that common dictionaries usually contain more information about a specified word, that WordNet does. Except for a simple information such as origin of the word, morphological data (genitive form, plural form), typical collocations, etc., every verb in a common dictionary can display its valency, which may represent quite a complicated structure. From that point of view it is much easier to store relations between literals in the other dictionary. Each word description then can contain an identification of a synset (given by ILI) which specifies which word meaning it belongs to.

Fig. 5 shows a VisDic definition for a simple common dictionary with a link to a WordNet database stored in the ENTRY.SYNSET tag. The format is followed by an example of two entries of this dictionary. Notice that all literal relations are stored in this dictionary instead in WordNet itself (especially ANTONYM relation, for example). The external tag ENTRY.SYNSET allows to work with the corresponding WordNet synset, as if it were included in the common dictionary. In the first synset it is linked via 06193747-n value, in the second one, the value of external synset is 05847495-n.

```
1 ENTRY                   1       1 N
  2 ID                    1       1 K
  2 HEAD                  1      -1 N
  2 PLURAL                0      -1 N
  2 IS_DERIVED_FROM       0      -1 L
  2 HAS_DERIVED           0      -1 R ENTRY.IS_DERIVED_FROM
  2 DERIVATION            0      -1 L
  2 PERTAINS_TO           0      -1 L
  2 IS_PERTAINED_TO       0      -1 R ENTRY.PERTAINS_TO
  2 ANTONYM               0      -1 L
  2 SYNSET                0       1 E SYNSET wn/en/wn_en
```

```
<ENTRY>
  <ID>00000001</ID>
  <HEAD>man</HEAD>
  <PLURAL>men</PLURAL>
  <ANTONYM>00000002</ANTONYM>
  <SYNSET>
    <POS>n</POS>
    <SYNONYM>
      <LITERAL>adult male
        <SENSE>1</SENSE>
      </LITERAL>
      <LITERAL>man
        <SENSE>4</SENSE>
      </LITERAL>
    </SYNONYM>
    <ILI>06193747-n</ILI>
    <HYPERONYM>05850734-n</HYPERONYM>
  </SYNSET>
</ENTRY>
```

```
<ENTRY>
  <ID>00000002</ID>
  <HEAD>woman</HEAD>
  <PLURAL>women</PLURAL>
  <ANTONYM>00000001</ANTONYM>
  <SYNSET>
    <POS>n</POS>
    <SYNONYM>
      <LITERAL>adult female
        <SENSE>1</SENSE>
      </LITERAL>
      <LITERAL>woman
        <SENSE>3</SENSE>
      </LITERAL>
    </SYNONYM>
    <ILI>06434591-n</ILI>
    <HYPERONYM>05847495-n</HYPERONYM>
  </SYNSET>
</ENTRY>
```

Fig. 5. Example of a common dictionary embedding the WordNet data

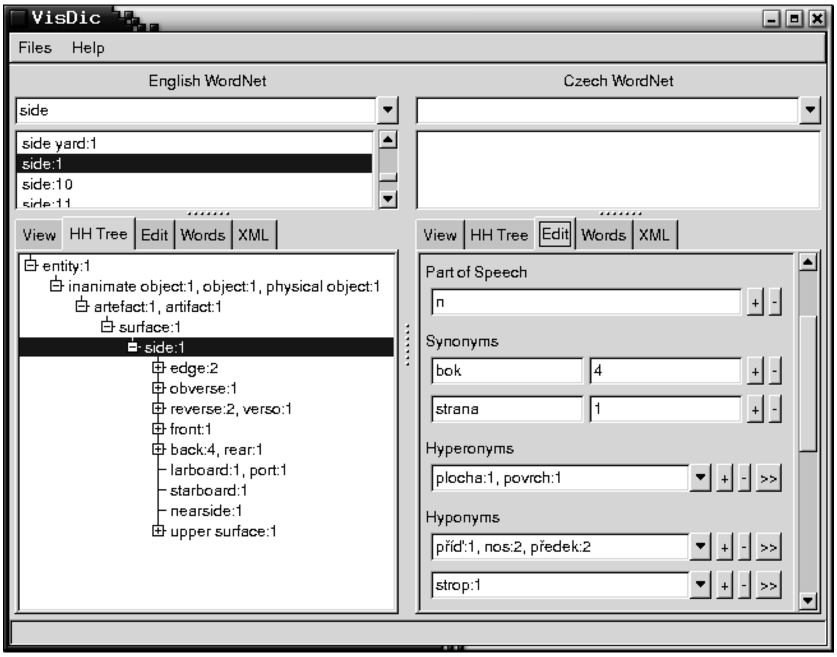
5. VisDic

VisDic is a program tool which allows to browse and edit common dictionaries, corpora and also databases like WordNet. All of these resources are based on elementary structures – common dictionaries consist of entries, while WordNet is made of synsets.

The user can view more dictionaries at the same time. Each has its own sub-window consisting of three parts. The topmost one is a query box. The middle one contains all found entries and the last displays a view of a specified entry. This window is represented by a graphical item called notebook which allows to view the entry in more ways. VisDic window with two active dictionaries can be seen in Fig. 6.

The query consists of an XML tag specification, = character and a value specification, e.g. if a user likes to find all the nouns in WordNet, he has to type *SYNSET.POS=n*. One of tags can be understood as the default one. Then the tag specification and = character can be omitted, e.g. if *SYNSET.SYNONYM.LITERAL* is defined as the default tag for WordNet database all the occurrences of a word form, say *side*, can be found by typing just *side*. Queries can be grouped by logical OR (*||*) or AND (*&&*), the value can be prefixed by ^ character, which means to find all entries beginning with the value phrase, or suffixed by *$* character, which means to find all entries ending with the value phrase.

The more complete description of VisDic can be found in (Pavelek, 2002).



side

side yard:1
side:1
side:10
side:11

entity:1
inanimate object:1, object:1, physical object:1
artefact:1, artifact:1
surface:1
side:1
edge:2
obverse:1
reverse:2, verso:1
front:1
back:4, rear:1
larboard:1, port:1
starboard:1
nearside:1
upper surface:1

Part of Speech: n

Synonyms: bok 4; strana 1

Hyperonyms: plocha:1, povrch:1

Hyponyms: příď:1, nos:2, předek:2; strop:1

Fig 6. VisDic

6. Conclusions

The suggested format fully corresponds to the XML format as it is used for the data representation. Although we do not use a proper DTD specification fulfilling the requirements of the standard DTD in XML, the presented definition is quite similar to it and though it does not use some features that XML offers in general we think that it is well suited not only for wordnets, but also for other lexical resources as well, such as explanatory dictionaries, bilingual dictionaries, dictionaries of synonyms, corpora, etc. The format used within VisDic tool enables a user to browse and edit easily any type of database stored in it.

The conclusion that can be drawn from this exercise is the following:

The standards can be arrived at either from top (this is not our case) or from the bottom which is the solution presented here. The experience seems to show that real standards develop from the practical use shared by many users. Then the modifications from the top can be applied and adopted if the users can agree upon them.

7. References

Louw M., 1998. *Polaris User's Guide*. Lernout & Hauspie, Antwerp, Belgium

Miller G.A., Beckwith R., Fellbaum Ch., Gross D., Miller K., 1993. *Introduction to Wordnet: An On-line Lexical Database*. Princeton University

Pavelek T., Pala K., 2002. *VisDic - A New Tool for WordNet Editing*. 1st International WordNet Conference, Mysore, India.

Vossen, P., 1999. *Final Report on EuroWordNet, CD ROM*. Amsterdam University

Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet

Dan Tufiş *, Dan Cristea♥

* RACAI-Romanian Academy
13, "13 Septembrie", Bucharest 5, Romania
tufis@racai.ro
♥University A.I. Cuza
16, Berthelot, Iaşi 6600, Romania
dcristea@infoiasi.ro

Abstract

The requirements in building a multilingual ontology of the EuroWordNet kind are frequently conflicting and if not considered in the first stages of the project, later harmonizing might be extremely difficult if possible at all. To ensure as early as possible usability, the incrementally developed lexical stock of each individual wordnet, should cover the most frequent vocabulary of the language. On the other hand, given that this is a multilingual lexical resource, special care should be addressed to the compatibility problems. Specifically, there are two main compatibility issues to be considered: there should be a cross-language conceptual coverage, meaning that each monolingual lexicon should globally deal with the same conceptual areas or domains and the interpretation of the defined relations should be the same in any monolingual ontology considered by the multilingual harmonized ontology. This is why, drawing as much as possible from the EuroWordNet lessons, we decided to address these issues at the very beginning phase of the BalkaNet project.

1. Introduction

BalkaNet [1] (Stamou et al, 2002) is an EC funded project (IST-2000-29388) that aims to develop in accordance with EuroWordNet philosophy a core multilingual resource for the following Balkan languages: Greek, Turkish, Romanian, Bulgarian, Czech and Serbian. As in EuroWordNet, the monolingual lexical ontologies are projected onto an interlingual set of concepts (ILI), the correspondences being established by means of complex equivalence relations (eq-synonymy, eq-near-synonymy, eq-has-hyperonym, eq-has-hypernym etc).

The requirements in building a multilingual ontology of the EuroWordNet kind are frequently conflicting (Rodriguez et al, 1998) and if not considered in the first stages of the project, later harmonizing might be extremely difficult if possible at all. To ensure as early as possible usability, the incrementally developed lexical stock of each individual wordnet, should cover the most frequent vocabulary of the language. On the other hand, given that this is a multilingual lexical resource, special care should be addressed to the compatibility problems. Specifically, there are two main compatibility issues to be considered: there should be a cross-language conceptual coverage, meaning that each monolingual lexicon should globally deal with the same conceptual areas or domains and the interpretation of the defined relations should be the same in any monolingual ontology considered by the multilingual harmonized ontology. This is why, drawing as much as possible from the EuroWordNet lessons, we decided to address these issues at the very beginning phase of the BalkaNet project.

The first part of the paper will address the approach we took for the selection of the initial lexical stock to be included into the Romanian core wordnet so that to observe multilingual design criteria and cross-language compatibility issues. The synsets (in two or more languages) that are mapped onto the same ILI concept are implicitly semantically linked. The nature of these cross-lingual semantic links, which we call *translational links*, depends on the links between the ILI concept and the synsets in the monolingual wordnets. One way to check consistency of the ILI projection of the individual wordnets is comparing the translation links with the translation equivalents licensed by a parallel corpus. This issue will be discussed in the second part of the paper.

2. An overview of the language resources

The Romanian wordnet started, as in the case of other languages in this project, from scratch. However, in order to ease the work and make the process as reliable as possible we built on various valuable language resources and several tools we developed for their exploitation. In the following there is a brief account of these building blocks, each of them being largely described elsewhere.

2.1. Corpora

Within the Multext-East and TELRI European projects (Erjavec et al. 1997), (Dimitrova et al., 1998), (Tufiş, Bruda, 1997), (Tufiş et al. 1997, 1998, 1999) there were created one 7-language heavily annotated parallel corpus based on Orwell's famous novel "1984" and one 25-language heavily annotated parallel corpus based on Plato's "The Republic". The annotation initially used was TEI compliant, but it was later on converted into CES (Ide, 1998). These are two relatively small corpora (about 110,000 tokens in each language) but given the accuracy of tagging and interlingual sentence alignment (hand validated) they were extremely useful for various applications ranging from building language models for morpho-syntactic tagging (Tufiş, 1999) and document classification (Tufiş et al., 2000) to automatic sense

[1] Further information can be obtained from the project's web site http://dblab.upatras.gr

discrimination (Erjavec et al., 2001). Besides the multilingual corpora we constructed two other much larger monolingual corpora: a literary corpus based on various novels (containing about 1,500,000 tokens) and a journalistic corpus (containing more than 100,000,000 tokens). Both corpora were automatically tokenized, tagged and lemmatized.

2.2. Lexicons and dictionaries

One delivery of the Multext-East project was a large wordform lexicon (more than 450,000 entries) containing triples <wordform, lemma, morpho-syntactic_code>. The encoding used in this lexicon is compliant with the Eagles recommendations for morpho-syntactic annotation and is largely documented in (Tufiş et al. 1997).

The reference dictionary we used for our analysis is The Explanatory Dictionary of Romanian (DEX,1996), work of the Romanian Academy Institute of Linguistics. This most authoritative lexicographic source for contemporary Romanian was partially digitized and converted into a lexical database (XML encoded) by RACAI under the European Project CONCEDE (Tufiş et al.1999). This core XML-dictionary has been extended to the full content of the printed dictionary by a follow-up project funded by Romanian Academy.

Another extremely useful lexical resource we relied on was the Romanian Dictionary of Synonyms-RDS (Seche, Seche 1997), which was transposed into electronic form by the NLP group at the University A.I. Cuza din Iaşi. The electronic form of RDS has been converted into an XML format so that the same query interface we developed for DEX works also with RDS.

From the multilingual parallel corpora mentioned before and using our translation equivalents extraction program (Tufiş, Barbu 2000, 2001a, 2001b) we constructed a bilingual Romanian English dictionary (also XML-encoded). This bilingual lexicon has been hand validated and extended with new entries from several public domain sources.

Finally, an extremely valuable resource was the ILI of the EuroWordNet, exported in XML format by means of the VisDic editor produced by the Masaryk University of Brno (Pavelek and Pala, 2002).

All these resources have been integrated by means of a series of tools developed for the purpose of the BALKANET project. They are user-friendly and allow for editing and mapping the Romanian synonymy series in RDS to the sense definitions in DEX and ILI records from EuroWordNet. The output of these tools is further subject to primary local consistency checks (such as detecting word sense appearing in more than one synset) and generated as an XML-encoded file appropriate for import in VisDic. We will provide a brief overview of these tools in Section 5.

3. Lexical stock selection

In order to ensure practical utility for the core wordnets to be delivered by the BALKANET project and to facilitate further extensions towards as large as possible coverage for the languages concerned, the project consortium decided to start the development process with a common set of concepts likely to be lexicalized in all the project languages. This special set of concepts, called *Base Concept* Set, was selected from the EuroWordNet interlingual index for reasons convincingly argued in (Vossen, 1998). The Base Concept Set contains 1310 concepts, each of them being attached a gloss and a Top Ontology Description (see Vossen, 1998). All project partners developed in a harmonized way the synsets in their languages corresponding to the Base Concepts. After this step, the monolingual wordnets will be further developed in a top-down approach starting with the synsets already mapped onto the Base Concepts.

Let us give a few definitions for some notions that will be used in the following.

When we place ourselves in a monolingual environment we speak about *senses*, *meanings* and *synsets.* A word has one or more *senses.* A sense refers to one *meaning.* In EuroWordNet the senses of a word are numbered according to their frequency and a sense of a lemma is denoted by appending the sense number to orthographic form of the lemma in case. A set of such numbered senses (eg. action2 activity1 activiteness1) referring to the same meaning is called a synset, which itself stands as a denotation of the common meaning of the senses in the synset. A meaning has a gloss that obviously applies for all senses in a corresponding synset.

When we want to abstract away from one language, we speak about the *concepts* referred to by the *word meanings*. So, we may speak about concepts with or without the reference to a specific language. Therefore, in trying to establish cross-lingual dependencies, via an interlingual index, it is convenient to refer to the entities used for this purpose as *concepts.* A concept is a language independent cognitive construct, which in EWN is always lexicalized at least in one language. A concept is further refined in terms of basic semantic distinctions (semantic features, sometimes referred to as semantic fields) so that one could speak about concept clustering along the basic semantic features.

According to these definitions we will use the term *Base Meaning* to refer to a basic (language specific) meaning in terms of which other word meanings can be defined and *which is directly mapped on a Base Concept.*

In EuroWordNet, and thus in BALKANET, ILI is defined as an unstructured collection of concepts represented by records of the form (<ILI-index> <ontological description> <gloss> {<domain>}). The initial ILI has been constructed from Wordnet1.5 and thus the gloss of each concept has been imported directly from the English synset referring to the meaning conceptualized in ILI.

According to the aims of the project regarding the interlingual coverage, language representativity, maximum usage of the core wordnet and scalability we started a series of quantitative analysis on a very large corpus made of several novels and a collection of journalistic texts, collected from the web. The corpus (containing more than 100 million words) was automatically tagged, lemmatized and the content words of interest (common nouns, verbs, adjectives and adverbs) were counted and sorted according to their frequency. We extracted this way, a list of more than 30,000 Romanian lemmas. Based on the frequency in the running texts, this list was divided into three parts, corresponding to the first 10,000 most frequent lemmas (I), the next most frequent 10,000 lemmas (II) and rest of the lemmas (III).

In deciding which is the most important subset of a lexical stock for a language, the frequency in running texts is considered by many lexicographers to be a very subjective criterion. Among the strongest arguments they would come with is the volume and representativity of the texts included into the corpus subject to the quantitative analysis. With more and more texts available on the net, the size of the data is not anymore a significant issue, but the representativity remains a systematic complain. The exact definition of what representative texts should be included into a corpus for quantitative data analysis is a long-standing debate and we won't get into this. Considering that our data consisted, almost entirely, of journalistic texts, the representativity issue could certainly be raised. The Frequency Dictionary of Romanian Words–FDRW (Julliand et all., 1965) published long time ago, based on a balanced corpus of 500,000 words of Romanian literature, legal texts, poetry and journalism contains a list of most frequent 5,000 lemmas. In spite of being quite contested, it is still used by many Romanian linguists as a reference. The comparison we made revealed that most of the 5000 words in FDRW were also in our list, although not with the same frequency ranges.

As frequency in running texts is a disputable criterion for deciding what words should be encoded into a core dictionary/thesaurus/ontology we considered that this criterion should be complemented with others, less controversial in the world of traditional lexicography.

Among the criteria one could find pleas for, we opted for two that we could easily turn into operational selectors. The one is the number of senses a headword would have in a reference dictionary. The second one is the number of word definitions that use the headword in case. A third criterion, not considered yet, might be the number of derivatives of a given headword (this last criterion is preferred by most Romanian etymologists).

In this phase of the BALKANET project we concentrated our attention to the Romanian nouns and the experimental data reported below refers to nouns. Since the technical procedures do not depend on the specific part of speech, the same would apply for verbs, adjectives and adverbs.

Considering only the first two frequency ranges described above (the first most 20,000 words in the journalistic corpus) we extracted from our Explanatory dictionary more than 8000 entries for nouns and nominal compounds (accounting for almost 35,000 senses) so that the definitional productiveness DP (the number of sense definitions a noun participates in) was at least 3. The list was sorted according to the definitional productivity.

Noun	**Definitional productivity**	**Number of definitions**	**FRECV$_{range}$**
acțiune	2279	13	I
persoană	1979	9	I
parte	1882	94	I
formă	1286	21	I
obiect	1204	16	I
fapt	1044	11	I
apă	743	29	I
• • •	• • •	• • •	• • •
rasism	3	1	II

Table 1: scoring the headword candidates

For all these nouns we extracted EN translations from our translation equivalence dictionary. The procedures for automatic extraction of translation equivalents from parallel corpora as well as the sense discrimination procedure are largely described in (Tufiş&Barbu, 2001a,b), (Erjavec et al, 2001). As the translation equivalents found by our extractor are limited by the available parallel corpora we have, provisions were made for automatic updating of the Ro-En dictionary with web resources.

All pairs containing an English word (or a synonym of it) in the English synsets corresponding to the base concepts were also associated with the corresponding top-ontology description. Practically for all English words corresponding to the base concepts there were found translations in our translation lexicon and these translations appeared in the upper top of our 8000-noun list. Those few EN nouns not translated in our lexicon were given manual translations. Because our translation equivalence lexicon is based on sense equivalence in context, transferring the ontological description from one EN word to its equivalent translation was considered to be a legitimate option. Thus, at the end of this step we collected a list of Romanian nouns associated with one or more English translations out of which at least one was present in the base concept list. Each such an association was further enriched with additional information extracted from other resources:

a) the RO word was attached with all its definitions extracted from the Explanatory Dictionary of Romanian;

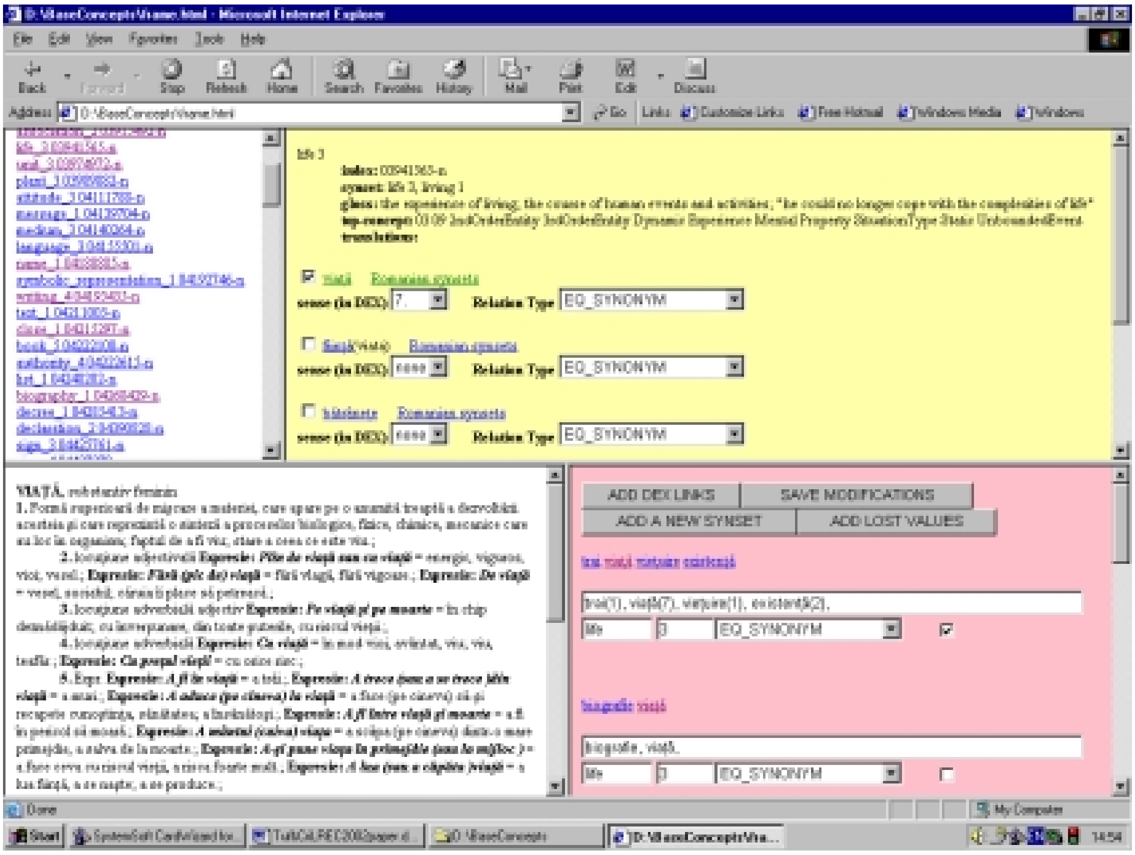
b) the EN word was attached with its entry in the WordNet1.5

The Romanian Dictionary of Synonyms (RDS), digitized and encoded as an ACCES database by University A.I. Cuza of Iaşi, was used to extract the synonymy series for the selected RO words. In RDS some members of the synonymy series are provided with usage information (old, regionalism, specific area of usage, domain, etc). Preliminary discussions lead to the idea to eliminate all the words marked as such (based on the assumption that we would like to construct a lexical stock for general use in contemporary Romanian). However, if later on this filtered out words (together with their usage information) would be necessary, their recovery was ensured. The synonymy series were taken as possible Romanian synsets and added to the RO-EN associations described above.

We have thus assembled the basic linguistic material that the lexicographer should use in making the decisions (linking) necessary for building the noun subset of the core Romanian wordnet. All this information is currently available in a java-based editor, showing in different frames, the following information (see figure 1):

- the list of the base concepts (upper-left frame), identified by the ILI record and an English word in the synset mapped on this concept (ex. *life_3_03941565-n*)
- the synset (life_3 living_1), its gloss and top-ontology description, possible translations and association boxes (right-upper frame)
- the numbered sense definitions from the Explanatory Dictionary of Romanian for the selected translation (left-lower frame);
- synonyms of the selected Romanian translation word (right-lower frame)

- pop-up menus for selecting the relevant sense numbers and the equivalence relation to the ILI concept.



life 3

index: 03941363-n
synset: life 3, living 1
gloss: the experience of living; the course of human events and activities; "he could no longer cope with the complexities of life"
top-concepts: 03 09 2ndOrderEntity 3rdOrderEntity Dynamic Experience Mental Property SituationType Static UnboundedEvent
translations:

☑ viață Romanian synsets
sense (in DEX): 7. **Relation Type** EQ_SYNONYM

☐ ființă(viață) Romanian synsets
sense (in DEX): 6666 **Relation Type** EQ_SYNONYM

☐ trăire(viață) Romanian synsets
sense (in DEX): 6666 **Relation Type** EQ_SYNONYM

VIAȚĂ, substantiv feminin
1. Formă superioară de mișcare a materiei, care apare pe o anumită treaptă a dezvoltării acesteia și care reprezintă o sinteză a proceselor biologice, fizice, chimice, mecanice care au loc în organism; faptul de a fi viu; stare a ceea ce este viu.;
2. locuțiune adjectivală **Expresie:** ***Plin de viață sau cu viață*** = energic, viguros, vioi, vesel.; **Expresie:** ***Fără (pic de) viață*** = fără viață, fără vigoare.; **Expresie:** ***De viață*** = vesel, sociabil, căruia îi place să petreacă.;
3. locuțiune adverbială adjectiv **Expresie:** ***Pe viață și pe moarte*** = în chip deznădăjduit, cu înverșunare, din toate puterile, cu riscul vieții.;
4. locuțiune adverbială **Expresie:** ***Cu viață*** = în mod viu, aievea, viu, teafăr.; **Expresie:** ***Cu prețul vieții*** = cu orice risc.;
5. Expr. **Expresie:** ***A fi în viață*** = a trăi.; **Expresie:** ***A trece (sau a se trece) din viață*** = a muri.; **Expresie:** ***A aduce (pe cineva) la viață*** = a face (pe cineva) să-și recapete cunoștința, sănătatea; a înviora.; **Expresie:** ***A fi între viață și moarte*** = a fi în pericol să moară.; **Expresie:** ***A scăpa (salva) viața*** = a scăpa (pe cineva) dintr-o mare primejdie, a salva de la moarte.; **Expresie:** ***A-și pune viața în primejdie (sau la mijloc)*** = a face ceva cu riscul vieții, a risca foarte mult.; **Expresie:** ***A lua (sau a căpăta) viață*** = a lua ființă, a se naște, a se produce.;

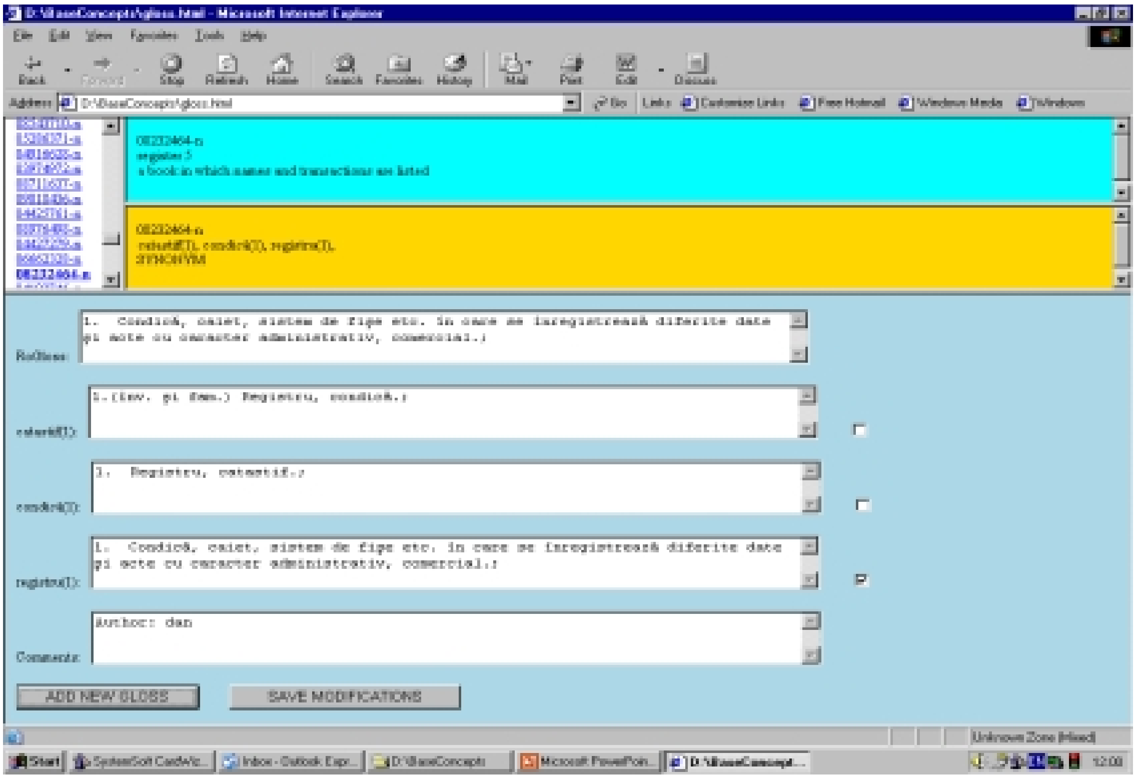
trai viață viețuire existență

trai(1), viață(7), viețuire(1), existență(2),
life 3 EQ_SYNONYM ☑

biografie viață

biografie, viață,
life 3 EQ_SYNONYM ☐

Figure 1: The editor for building synsets for the base meanings



08232464-n
register 3
a book in which names and transactions are listed

08232464-n
catastif(1), condică(1), registru(1),
SYNONYM

RoGloss: 1. Condică, caiet, sistem de fișe etc. în care se înregistrează diferite date și acte cu caracter administrativ, comercial.;

catastif(1): 1.(Inv. și fam.) Registru, condică.; ☐

condică(1): 1. Registru, catastif.; ☐

registru(1): 1. Condică, caiet, sistem de fișe etc. în care se înregistrează diferite date și acte cu caracter administrativ, comercial.; ☑

Comments: Author: dan

D:\BaseConcepts\gloss.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address D:\BaseConcepts\gloss.html
Go Links Customize Links Free Hotmail Windows Media Windows
08232464-n
ADD NEW GLOSS
SAVE MODIFICATIONS
Unknown Zone (Mixed)
Start
SystemSoft CardWiz...
Inbox - Outlook Expr...
D:\BaseConcepts
Microsoft PowerPoin...
D:\BaseConcept...
12:00

Figure 2: The editor for gloss assignment

The editor has been instantiated into 10 differently populated copies, each containing a different set of base concepts. Each incarnation of the editor has been given to a different expert who was in charge of building his/her set of Romanian synsets and map them onto the appropriate base concepts. When this building phase was finished we performed a few simple error-checking such as:

- all literals appearing in a synset should have attached a sense number
- no sense (literal and sense number) should appear in two or more synsets
- each synset should have an equivalence relation to a unique base concept.

Once the synsets were constructed and mapped onto base concepts, the second phase was to add a Romanian gloss to each Romanian synset. In the vast majority of cases, the definitions extracted from DEX corresponding to the senses in a synset were different in wording so, the lexicographers had to chose the best definition, closest to the definition of the corresponding base concept. The Figure 2 shows that the base concept 08232464-n corresponding to the 5th sense of the English word *register* (a book in which names and transactions are listed) corresponds in Romanian to the synset (catastif_1 condică_1 registru_1). The selected senses for the three Romanian words have in DEX different definitions. By checking the box to the right of the third definition (lower frame in Figure 2) the lexicographer decided that the definition given to *registru_1* is the one to be attached to the synset.

It is worth mentioning that during the gloss assignment phase it became apparent that several synsets were not correct, requiring modifications. In some cases, the Romanian Explanatory Dictionary includes under the same definition two senses that are differentiated in ILI as two distinct concepts. In such cases, the general strategy was to split the Romanian definition and attach the relevant part as a gloss.

4. A proposal for cross-lingual validation of the ILI mapping

As we said before, one of the main objectives of the BALKANET project (which adopted a merge model approach) is to ensure as much as possible overlap between the concepts lexicalized in the concerned languages. A significant overlap may be hampered either by conceptually different lexical stocks for the different languages or by inconsistent projection of the monolingual concepts onto the ILI concepts. In order to ensure conceptual similarity for the lexical stocks across various languages, the development of the monolingual ontologies started in two different, but convergent ways: the minimalist one was to provide direct translations of the EuroWordNet Base Concept Set; the second way (language-centric) was to produce a ranked list of most important (according to prescribed lexical criteria) words in each language and to include in the monolingual wordnets at least those words, the meanings of which would cover the Base Concept Set. Irrespective of the approach taken towards ensuring lexical stock similarity across languages, we had to consider means for automatic check of the correctness of the mapping of the monolingual synsets over the ILI concepts. To this end we will describe in some details a proposal for an automatic consistency checking.

Our approached is based on the notion of translation equivalence over bitexts, on bilingual lexicons automatically extracted from parallel corpora (Tufiş, Barbu, 2001 a,b) and on sense disambiguation (Erjavec et al., 2001).

The parallel corpus we used in our experiments is the "1984", based on Orwell's famous novel, developed in the MULTEXT-EAST project, further cleaned up in the TELRI and CONCEDE projects. The corpus contains professional translations of the original novel in 6 languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene), all aligned at the sentence level to the English original. Each monolingual part of this 7-language parallel corpus is segmented, tagged and lemmatized and also carefully hand validated.

From the 6 (integral) bitexts (CEE language texts aligned to the EN original) there were extracted bilingual lexicons (XX-EN, with XX one of the six CEE languages) and furthermore a 7-languages lexicon with EN as a hub. By removing all the non 1-1 alignments in the bitexts and using the EN sentence Ids as anchors, a partial (about 92% of the whole text) 7-lingual 1-1 alignment (EN-BG-CZ-EE-HU-RO-SI) was computed. The 7-language aligned corpus allows for extracting any of the 21 possible (partial) bitexts. A number of 104 nouns appearing in the English part of the multilingual corpus (altogether 3316 instances were hand annotated and used as a gold standard for our sense clustering algorithm (Erjavec, 2001).

As BG, CZ and RO are languages of the BalkaNet project from the present data, our methodology could be used for checking the ILI-mapping consistency for any of the RO-EN, RO-CZ, RO-BG, EN-CZ, CZ-BG and BG-EN pairs of wordnets. In the current phase of the project we are able to consider only the interlingual mapping of the base concepts. Let us generically denote the language pairs subject to checking as XX-YY. The basic methodology is as follows:

1) From the XX-YY bitexts we extracted the XX-YY lexicon (http://www.racai.ro/~tufis/BilingualLexicons/AutomaticallyExtractedBilingualLexicons.html). The bilingual lexicon contains not only the translation pairs but also, for each entry the aligned sentences that licensed the translation equivalence relation. This lexicon is purged so that it contains only words that have (in the respective monolingual wordnets) at least one sense mapped on a base concept set. Put it otherwise, any pair (*W_{XX} translated as W_{YY}*) of the purged lexicon has the property that W_{XX} or W_{YY} or both have at least one sense in the language-specific base meaning set.

2) Let it be (W_{XX} W_{YY}) a translation equivalent. Let us denote with S_{WXX} the synsets in language XX containing the W_{XX} word (actually one sense of it) and S_{WYY} the synsets in language YY containing the W_{YY} word (actually one sense of it). Starting in the XX monolingual wordnet from the synsets in S_{WXX}, via ILI, one ends in the YY monolingual wordnet with the XX-synsets having translation links to YY-synsets. Let us call this set as S'_{WYY}. S_{WYY} and S'_{WYY} should have at least one synset in common. Please note that if the intersection of the two sets of synsets is non-empty, the described procedure ensures semantic tagging of the (W_{XX} W_{YY}) pair with one or more ILI-concept tags. If the intersection contains exactly one synset, its corresponding ILI record-number

could be used to semantically tag both W_{XX} and W_{YY}. With intersection containing more synsets, we still are able to reduce the semantic ambiguity of the considered words. In case the intersection is empty, we might have one of the following possible explanations:

2.1) (W_{XX} W_{YY}) is not a valid translation pair; by checking the sentences that licensed the extraction of this translation pair one could confirm or refute this possibility; please note that an error here might be due to the extraction algorithm or to a problematic human translation (for instance it is not uncommon that even professional translators would sometimes translate one word by a non-eq-synonym for various reasons like contextual semantic gaps or stylistic preferences)

2.2) (W_{XX} W_{YY}) is a valid translation pair and the two words share a meaning assigned to a concept which is not in the base concept set.

2.3) the interlingual mapping of the W_{XX} and W_{YY} is “wrong”; being “wrong” might be a real mapping error in the XX or YY language (or in both) or it might be motivated by a lexical gap in one of the languages concerned (or both); the lexicographer might have overcome the lexical gap by using a complex equivalence relation (not the eq-synonym); in the second case, one might get insights on possible concept clustering at the ILI level (creating so-called *soft-concepts*).

We claim that this procedure allows us to estimate both the cross-lingual coverage and the correctness of the interlingual mapping of the two considered monolingual wordnets. The procedure allows not only for estimation but also for pinpointing the incomplete or missing synsets as well as inconsistencies in mapping the synsets onto ILI concepts and gives hints on soft-concept clustering.

4.1. Condiments, spices, sauces and other ingredients

Let us consider the fragments of the Ro-Wordnet and WN1.5 shown in the Figure 3. The arrows represent hyponymy relations in the two wordnets. The gray heavy lines represent translational links between the synsets in the two languages, meaning that the respective synsets are mapped onto the same ILI concept. The heavy dashed line represents a translational link that is reported as wrong during the cross-validation of the two wordnets. The reason for this comes from the violation of what we called the *hierarchy preservation principle*. The inconsistency is signaled because in language RO the hierarchical relations (hyponym) between ^{M}*mirodenie*$_{RO}$ H ^{M}*condiment*$_{RO}$ as well as ^{M}*ketchup*$_{RO}$ H ^{M}*sos*$_{RO}$ are not verified in language EN by the equivalent pair meanings (^{M}*spice*$_{EN}$ ^{M}*condiment*$_{EN}$) and (^{M}*ketchup*$_{EN}$ - ^{M}*sauce*$_{EN}$)(in EN they are sisters). If the structuring in WN1.5 is taken to be the Truth, this example shows that *the hierarchy preservation principle* is not true. On the other hand, if it would be reasonable to consider that WN1.5 is amendable (for instance making ^{M}*mustard*$_{EN}$ and ^{M}*ketchup*$_{EN}$ direct hyponyms of ^{M}*sauce*$_{EN}$) then the *hierarchy preservation principle* might be a very powerful consistency check.

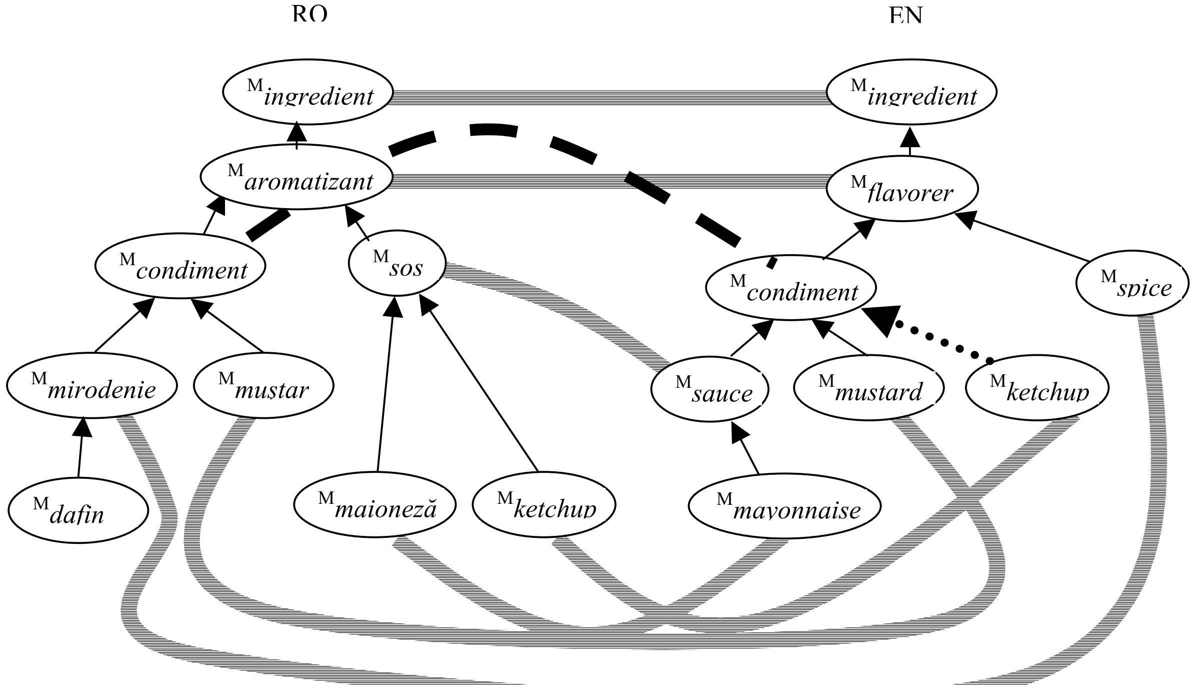


Figure 3: Translational links and consistency checks

5. Conclusions and further work

The approach on consistency checking based on translation equivalents in multilingual parallel corpora has some methodological similarity with (Resnik et al., 1999) on the multilingual corpus built up from many translations of the Bible. Speaking about useful sense distinctions (for machine translation for instance) Resnik (personal communication) identifies *strong sense distinctions* of one word in a source language as those that are lexicalized as different words in the target languages. When some senses carried by a source word are found in a target word the distinction between them is called a *light sense* distinction. In the area of machine translation trying to disambiguate among light distinctions is not a very productive enterprise and therefore being able to identify, for a given pair of languages, which are the strong/light sense distinction might be extremely useful for machine translation. Our approach could be used to enhance the strong/light dichotomy with a third dimension: *fuzzy sense* distinction. This term is strongly related to that of *soft concept* used in EuroWordNet for clustering different ILI concepts that are lexicalized in two or more languages by words considered to be legitimate translations of one another.

In the next phase of the project, in order to extend the monolingual Romanian wordnet up to the level of the promised size, our strategy will be language-centric meaning that the new entries will be the top ranked words selected from our noun/verb/adjective/adverb lists sorted as described in the section 3.

6. References

Bloksma L., Diez-Orzas and Vossen P. (1996) The User Requirements and Functional Specification of the EuroWordNet-project *EWN-deliverable D.001*, LE-4003

DEX (1996). Coteanu, I., Seche, L., Seche, M. (coord.). Dicționarul Explicativ al Limbii Române, Ediția a II-a, *Univers Enciclopedic*, Bucureşti, 1996

Erjavec T., Ide N., Tufiş D.(1997) Encoding and Parallel Alignment of Linguistic Corpora in Six Central and Eastern European Languages" in Michael Levison (ed) *Proceedings of the Joint ACH/ALL Conference* Queen's University, Kingston, Ontario, June 1997 (also on http://www.qucis.queensu.ca/achallc97)

Erjavec T., Ide N., Tufiş, D.(2001) *Automatic Sense Tagging Using Parallel Corpora*, in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 27-29 November, pp. 212-219, 2001

Ide, N. (1998) *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora* First International Language Resources and Evaluation Conference, Granada, Spain. See also http://www.cs.vassar.edu/CES/.

Julliand, A., Edwards P.M.H, Julliand I. (1965). The Frequency Dictionary of Rumanian Words. *Mouton & CO.,* London-The Hague-Paris, 1965

Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. (1990) "Introduction to WordNet: An On-Line Lexical Database" 1990 In International Journal of Lexicography, Vol. 3, No. 4 (winter 1990), pp. 235-244

Pavelek T., Pala K. (2002) *VisDic : A new Tool for WordNet Editing* in Proceedings of the 1st International Wordnet Conference, Mysore, January 21-25, 2002

Resnik, P. (1999) Disambiguating Noun Groupings with Respect to WordNet Senses, in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora.* Kluwer Academic Publishers, 1999, pp. 77-98.

Resnik, P., Broman Olsen M., Diab M.(1999) The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues', *Computers and the Humanities*, 33(1-2), pp. 129-153, 1999.

Resnik P., Yarowsky D. (2000) Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation, *Natural Language Engineering* 5(2), pp. 113-133.

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A.(1998) The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In Piek Vossen (ed.) *EuroWordNet: A Multilingual database with lexical semantic networks*, Computers and Humanities*,* Vol. 32, Nos 2-3, 1998

Seche L., Seche M.(1997) *Dicționarul de sinonime al limbii române.* Univers Enciclopedic, Bucureşti, 1997

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., Grigoriadou M..(1997) BALKANET A Multilingual Semantic Network for the Balkan Languages, in *Proceedings of the International Wordnet Conference*, Mysore, India, 21-25 January 2002

Tufiş D., Şt. Bruda (1997) Structure Markup in CES and Preliminary Statistics on Romanian Translation of Plato's "Republica", *Proceedings of International Seminar on Encoding,* Ljubliana, February, 1997, also in *TELRI News*, nr. 5, May, 1997.

Tufiş, D. Tiered Tagging and Combined Classifiers In F. Jelinek, E. Nöth (eds) *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, Springer, 1999

Tufiş D., Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. (1997). Corpora and Corpus-Based Morpho-Lexical Processing, in Tufiş D., P. Andersen (eds.) *Recent Advances in Romanian Language Technology*, Editura Academiei, 1997.

Tufiş, D., Rotariu, G., Barbu, A.M. (1999) TEI-Encoding of a Core Explanatory Dictionary of Romanian. In Kiefer, F. and Pajzs J. (eds.) *Papers in Computational Lexicography*, Hungarian Academy of Sciences, 1999, pp. 219-228

Tufiş D., Popescu C., Roşu R.: Automatic classification of documents by random sampling in *Proceeding of the Romanian Academy*, Series A, vol 1, no. 2, p. 18-28, 2000

Tufiş, D. (2000). Blurring the distinction between machine readable dictionaries and lexical databases. R*esearch Report, RACAI-RR56*, 1999

Tufiş, D. (2001) Romanian wordnet of BALKANET: selecting the lexical stock. R*esearch Report, RACAI-RR68*, October 2001

Tufiş, D, Cristea D. (2001) *Methodological issues in selecting the candidate concepts to be included into the Romanian Wordnet.* Progress Report on BALKANET project. November 2001

Tufiş D., Barbu A.M.(2001a) *Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries*, in International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453-8245, Vol.4, No.3-4, 2001, pp.325-352

Tufiş D., Barbu A.M.(2001b) *Extracting multilingual lexicons from parallel corpora*, in Proceedings of the ACH-ALLC conference, New York, 12-17 June, 2001.

Vossen P. (ed.) (1998) "A Multilingual Database with Lexical Networks", Kluwer Academic Publishers, Dordrecht

Semi-automatic Development of the Hungarian WordNet

Gábor Prószéky*, Márton Miháltz**

*MorphoLogic, Késmárki u. 8, H-1118 Budapest, Hungary
proszeky@morphologic.hu
**Eötvös Lóránd University, Institute of Informatics, Budapest, Hungary
mmarcy@inf.elte.hu

Abstract

Construction of the Hungarian WordNet began in 2000. The project presently focuses on the nominal part. Our principal approach is to use Princeton WordNet as the basic structure, to which Hungarian nouns are attached. We are applying two methods to accomplish this: manual disambiguation for the more abstract levels, and automatic methods, including heuristics developed by earlier projects, in order to attach the remaining more specific senses. Results from these methods are integrated into a core structure, which will be enriched using further electronic linguistic resources.

1. Introduction

The construction of the Hungarian WordNet has started from scratch, unlike many components of the EuroWordNet project (Vossen, 1999), whose creators could rely on already existing lexical resources (Kunze et al., 1998). We employed the initial hypothesis that nominal hierarchies in English and Hungarian should be similar, at least for certain domains. This enabled us to attach Hungarian nominal entries of a Hungarian-English bilingual dictionary to Princeton WordNet 1.6 (WN) synsets. In this way, the English nominal hierarchy of WN serves as a skeleton structure to support the construction of the core Hungarian nominal WordNet. This approach was also used in the initial stage of the construction of the Spanish and Catalan WordNets (Farreres et al., 1998; Atserias et al., 1997). Furthermore, examining the Hungarian nominal taxonomies extracted from a Hungarian monolingual dictionary, we have found that hierarchies for the specific nominal domains (nouns denoting objects) tend to be similar to those found in WordNet.

Linking Hungarian words to WN synsets is accomplished in two ways. First, a software environment has been created to support the manual disambiguation of Hungarian nouns against WordNet. This is a top-down procedure advancing from the abstract to more specific levels in the WordNet hypernym structure, resulting in the manual construction of the more abstract levels of the core Hungarian WordNet.

Secondly, various heuristics, mostly developed in earlier projects, are applied to produce sets of candidate links between Hungarian nouns and WN synsets automatically. These methods rely on information found in the bilingual and monolingual dictionaries, plus the information already available from the parallel manual disambiguation procedure. Finally, results of all the different methods are manually checked and integrated.

Attaching further nominal entries from a larger bilingual dictionary, a thesaurus, and entries and definitions (serving as glosses) from a monolingual dictionary will enrich the resulting skeleton Hungarian WordNet structure, connected to the English WN.

In the following section, we will give a review of the electronic resources we use in our work. Section 3 gives an overview of the various automatic and manual methods used in the project. Integration of the information from different sources and the possibilities for further extensions are also discussed. Finally, our conclusions are provided in Section 4.

2. Acquiring taxonomies from various dictionaries

We have several electronic resources at our disposal: English-Hungarian bilingual dictionaries, a monolingual Hungarian explanatory dictionary, a Hungarian Thesaurus, and, of course, WordNet 1.6. MorphoLogic's English-Hungarian bilingual electronic dictionary contains entries for 17,801 Hungarian nouns with 12,440 English translations included in WordNet. The dictionary has been converted into a database of English-Hungarian word pairs with symmetrical translation relations (Prószéky et al., 2001). The entries of the Hungarian side constitute the basic set used for the various attachment procedures (see: Section 3).

A significantly enlarged version of the *English-Hungarian Dictionary* (Országh–Magay, 2001) will be used for further improvement of the Hungarian WN structure. It contains over 150,000 Hungarian entries, with English translations covering more than 80% of WordNet's entries. An electronic version of the Hungarian explanatory dictionary *Magyar Értelmező Kéziszótár* (ÉKSz) (Juhász et al., 1972) has been converted into XML format. This dictionary contains 42,942 nominal entries, corresponding to 64,146 definitions. 31,023 of them are annotated with usage codes, representing either the semantic domain (sport, medicine, science, religion etc.), or the language usage (technical, slang, vulgar, intimate, etc.). Through the smaller bilingual dictionary, 10,507 headwords have English translations in WN. We also have at our disposal a Hungarian electronic thesaurus. The *Magyar Szókincstár* contains 25,500 entries with synonyms and 14,400 entries with antonyms. Entries are linked to separate sets of synonyms for various senses. Most of the synonym and antonym words are annotated with language usage labels.

To help the construction of the Hungarian nominal WN, information is acquired from the monolingual dictionary in several ways. First, programs were developed to parse each dictionary definition and extract semantic information. In 83% of all the definitions, genus words were identified, which can be accounted for as hypernym approximations of the corresponding headwords. For example, the following ÉKSz entry will tell us that the *koala* is a kind of *mammal*:

> **koala:** *marsupial* ***mammal*** *resembling a bear, native in Australia*

In about 1,700 cases, the identified genus word was either a group noun, or a word denoting a "part" relationship. Let us consider as an example the ÉKSz entries for *alphabet* and *face:*

> **alphabet:** *The* ***set*** *of letters used for…*
>
> **face**: *The* ***part*** *of the head that…*

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

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

We are experimenting with several heuristics, relying on the work by Rigau et al. (1997) and Copestake (1990) to achieve an automated process of genus word disambiguation. About 70% of the genus terms are monosemous in the monolingual dictionary. In these cases the hyponym senses are attached to them directly.

Another heuristic utilizes the usage codes available for about 30% of the candidate senses Semantic codes, if available, can be tested for compatibility between the hyponym and the candidate hypernym senses. The pragmatic codes are also put to use: senses annotated as slang, vulgar etc. are more unlikely to be used as genus terms.

A third heuristic assigns the first sense occurring in an entry, relying on the fact that senses are ordered by usage frequency, and the most used senses are more likely to be used as hypernyms.

A fourth heuristic tries to measure semantic similarity among definitions by means of determining the number of lemmas shared by both definitions.

A fifth heuristic will rely on the conceptual distance formula, which measures semantic similarity between concepts using WordNet as a hierarchical knowledge base (Rigau et al., 1997). Application of the conceptual distance formula is discussed in more detail in Section 3.2.2.

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

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

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

On the other hand, these facts have encouraged us to start linking Hungarian nouns manually, starting from the topmost WN levels, and to apply automatic linking procedures for the more specific senses.

3. Manual and semi-automatic procedures

We are using both manual and semi-automatic techniques to achieve the task of linking Hungarian nouns to the WN synsets. The manual methods provide a framework of top-down construction of the Hungarian nominal WordNet. The automatic methods rely on the bilingual and monolingual dictionaries, and on the extracted semantic information, applying heuristics developed for the construction of the Spanish and Catalan WordNet (Farreres et al., 1998; Atserias et al., 1997). We chose to test these methods because the resources available to the Spanish and Catalan Research Group are closest to our available resources, considering the participants in the EuroWordNet project (Vossen et al., 1999).

The result of these methods will be evaluated manually, based on random samples. Then all the possible intersections of the sets of results produced by the different methods will also be evaluated, and only the results obtained by the combination that produces the highest accuracy will be considered. We follow this approach, described by Atserias et al (1997), in order to ensure the precision of the core Hungarian WordNet structure.

3.1. Manual disambiguation with the help of the web

A set of Internet-based software tools has been developed for manual disambiguation of the Hungarian nominal entries against WN. The use of the Internet makes it possible for our contributing experts to work independently.

For the users, the system offers a web page, over which the expert can answer questions provided by the central server maintaining the database. (Figure 1) xperts are exposed to dialog boxes: if the word in question does mean the concept outlined below by English synonyms and a definition, then the human expert is supposed to press the Yes button (Nagy, 2001).

3.2. Semi-automatic methods based on heuristics

There are three kinds of automatic linking methods, each relying on different kinds of resources.

The first group of heuristics relies on information found in the bilingual dictionary and the structure of WN, while the second type relies on the genus information extracted from the monolingual dictionary. These constitute heuristics described by Atserias et al. (1997), plus a technique of our own.

The third method relies on the links already produced by the manual linking procedure and the taxonomy acquired from the monolingual dictionary.

3.2.1. Methods relying on bilingual dictionaries

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

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

3.2.2. Methods relying on monolingual dictionaries

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

The second type of our automatic methods that utilize the monolingual dictionary relies on the extracted genus information (see Section 2). Following Atserias et al. (1997), we are applying the *Conceptual Distance formula* for the English translations of each headword-genus, or headword-holonym word pair we identified in the dictionary. The Conceptual Distance formula, introduced by Agirre et al. (1994), selects those two closest concepts in WN which represent the two input words. In the case of headword-genus pairs, the hypernym structure of WN is used as a semantic network for the heuristic, while for the ÉKSz headwords with holonym/meronym word approximations, the structures determined by WN's holonym links are used.

The application of the Conceptual Distance formula not only produces candidate links for the Hungarian words, but can also be used as a heuristic in the sense disambiguation of the Hungarian genus words, thus contributing to the construction of the Hungarian nominal taxonomy (Rigau et al., 1997).

3.2.3. Using information from the manual disambiguation procedure

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

For example, let us suppose that the Hungarian word *állat* (`animal') has already been linked (either manually or automatically) to the WordNet synset {*animal, animate being, beast, brute, creature, fauna*}. *Állat* has 3 different senses in the Hungarian taxonomy, one of which has a hyponym pointer to (a sense of) the word *ló* ('horse'). The word *ló* has 3 English translations in the bilingual dictionary, which belong to 8 different synsets in WordNet. In order to determine which of those 8 synsets should *ló* be linked to, Conceptual Distance (see Section 3.2.2) is calculated between {*animal, animate being,…*} and the 8 candidate synsets. The candidate synset {*horse*, *equus caballus}* will show the smallest distance from the hypernym synset {*animal, animate being,…*}, thus, *ló* (with the sense determined by the hypernym *állat*) can be linked to {*horse*, *equus caballus*} (Figure 2).

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

3.3. Further steps

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

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

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

4. Conclusion

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

5. References

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

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

Copestake, A., T. Briscoe, P. Vossen, A. Ageno, I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou, 1994. Acquisition of Lexical Translation Relations from MRDs. In *Journal of Machine Translation*, 3.

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

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

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

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

Kunze, C., A. Wagner, D. Dutoit, L. Catherin, K. Pala, P. Ševeček, K. Vider, L. Paldre, H. Orav, H. Oim. 1998. *First WNs for BCs in French, German, Czech and Estonian*. EuroWordNet Deliverable 2D007.

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

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

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

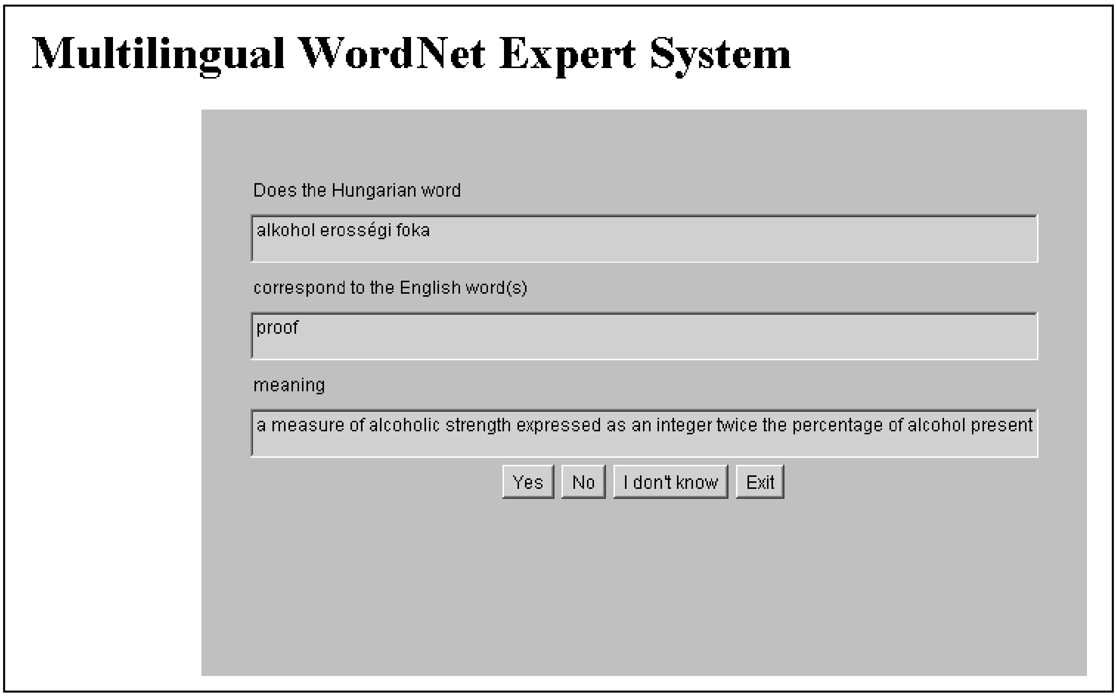
Pala, K., and P. Ševeček, 1999. *The Czech WordNet.* EuroWordNet (LE-8328) Deliverable 2D014

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

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

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

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



Multilingual WordNet Expert System

Does the Hungarian word

alkohol erosségi foka

correspond to the English word(s)

proof

meaning

a measure of alcoholic strength expressed as an integer twice the percentage of alcohol present

Figure 1. Disambiguation dialogue

```
Synonyms/Hypernyms (Ordered by Frequency) of noun ló
1 sense of ló
Sense 1
Equus caballus, ló                                                 (horse)
      => emlős                                                     (mammal)
          => állat                                                 (animal)
              => valami                                            (entity)

Hyponyms of noun ló
1 sense of ló
Sense 1
Equus caballus, ló                                                 (horse)
      => harci mén                                                 (steed)
      => amerikai félvad ló, musztáng                              (mustang)
      => versenyló                                                 (racehorse)
```

Figure 2. Sample hypernymy/hyponymy hierarchy

Viewing Semantic Networks as Hypermedia

Dimitris Avramidis*†, Maria Kyriakopoulou*†, Giorgos Kourousias†,
Sofia Stamou*†, Manolis Tzagarakis†

*Computer Engineering and Informatics Department
University of Patras
GR-265 00, Rion Patras, Greece
{avramidi, kyriakop, stamou}@ceid.upatras.gr

†*Research Academic* Computer Technology Institute
Riga Feraiou 61,
GR-262 21, Patras, Greece
{avramidi, kyriakop, stamou, gk, tzagara}@cti.gr

Abstract

The analogy of a semantic network to hypertext has long been recognized, and a semantic network has been considered as a logical model of hypertext – especially for those hypertexts with typed nodes and links. Moreover, wordnets form the most representative type of semantic networks in the field of Natural Language Processing and semantics in particular. It is obvious that hypertext and wordnets share many common points regarding their fundamental principles and the objectives towards which they both aim. In particular, they are both targeted towards capturing relations that possibly exist between objects and thus providing information of the underlying objects via various types of links used for describing the relations. In this respect, we strongly believe that if semantic networks are viewed beyond strictly linguistically constraints and applications, the results could only be beneficial.

1. Introduction

Hypertext[1] has always been closely related to the idea of freedom to associate, making it to be considered as an alternative means of structuring information. This new promising field provides its users (namely, authors and readers) with effective ways of presenting and exploring information. For authors, hypertext systems offer a high degree of flexibility for connecting pieces of information and presenting it as an assembled collection in an information network. For readers, hypertext provides tools for navigating in these information networks and for exploring them freely. Therefore, hypertext can be a precious dialogic means, facilitating the organization of information according to the user needs.

On the other hand semantic networks form a highly structured linguistic resource enabling a flexible navigation through the lexical items of a language. Wordnet forms a kind of conventional dictionary where semantic information of the terms it contains is represented. The main structural entities of wordnets are language internal relations through which words are linked based on their semantic properties. The main contribution of wordnets in lexicography is the systematic patterns and relations that exist among the meanings that words can be used to express. In this respect wordnets as a particular type of semantic networks resemble much hypermedia as far as the structural organization of information is concerned.

The paper is organized in the following way. Section 2 provides a brief overview of structure in semantic networks. In section 3, we reason about the ability of hypertext to structure information. Section 4 focuses on the similarities that hypertext and wordnets share, claiming that semantic networks can be viewed as hypertext. Finally, section 5 refers to the benefits that these two research areas may have if they are seen as a whole.

2. Structure in Semantic Networks

Wordnets form the most representative type of semantic networks in the field of Natural Language Processing and semantics in particular. Motivated by theories of human knowledge organization, wordnet emerged as a highly structured language repository, where words are defined relatively to each other. Unlike machine-readable dictionaries and lexica in book format, wordnet makes the commonly accepted distinction between conceptual-semantic relations, which link concepts and lexical relations, which link words (Evens, 1988). Thus, despite their resemblance to typical thesauri, wordnets in general clearly separate the conceptual and the lexical levels of language, and such a distinction is reflected via semantic-conceptual and lexical relations that hold among synsets and words respectively. Wordnets form semantic dictionaries that are designed as networks, partly because representing words and concepts as an interrelated system seems to be consistent with evidence for the way speakers organize their mental lexicons (Miller, 1998; Kay, 1989).

Wordnets' hierarchical structure allows a searcher to access information stored in lexical chains along more than one path, semantics being among them. Conceptual structures are modelled as a hierarchical network enabling a graphical representation of the lexicalized concepts when the latter are denominated by words (Priss, 1998). The theoretical analysis shows dependencies among semantic rela-

[1] Initially, hypertext dealt only with the manipulation of text. Nowadays, one can shape information structures containing pictures, video, sound, etc. Hypermedia – a contraction of the words *Hypertext* and *Multimedia* – is a name invented to stress this change of emphasis.

tions, such as inheritance of relations from sub-concepts to super-concepts. Therefore, related senses grouped together under the same lexical chain form preliminary conceptual clusters. Words belonging to the same lexical chain are connected via language internal relations, each one denoting the type of relation that holds among the underlying word meanings. Some of the language relations are bi-directional in the sense that if a link holds between term A and B then a link also holds between term B and term A. However, bi-directionality of the relations strongly depends on the language particularities and semantic properties of the underlying word meanings.

In order to account for particularities in lexicalized concepts, tags are assigned to each lexical relation denoting specialized semantic characteristics of a word's meaning. Tags can be viewed as a means of semantic constraints posed upon semantic relations that link word meanings rather than word forms. Moreover, tags provide information about which of the semantic properties represented in a lexical chain are inherited to its components. In this respect, words represent an atomic and unbiased level of individuality that becomes meaningful via anchoring of semantic relations. As Hasan (Hasan, 1984) pointed out, any word in a chain can be related to multiple other words in that chain. All lexical relations form a graph where cycles are disallowed since after all they contribute very little of any new information.

Summarizing, the structure of lexical data within wordnets is what differentiates the latter from traditional lexicographic aids (both dictionaries and thesauri). The motivation behind construction semantic networks in the form of a graph relies on the fact that lexical data becomes meaningful only via predefined linguistics structures. Navigation through the content of wordnets becomes feasible via language internal relations, which form the main notion around which structure is defined.

3. Hypermedia Principles of Structure

The term of hypertext cannot be explicitly defined since one can approach it by different directions. More specifically, there are those who claim that hypertext can be viewed as an interaction paradigm, referring to the manipulation of "pointing at a link and clicking it" in order to follow it. Additionally, there are others maintaining that "hypertext deals with the organization of information", regarding not only data but also structure as first-class user abstractions. Finally, there is another user group that considers "structure more important than data", making hypertext more structure-based technology than data-dependent.

Adopting the "primacy of structure over data" (Nürnberg et al., 1997), hypertext can be seen as a technology well suited to exploring different kinds of representational structures (Marshall, 1987). Viewing different parts of information as objects, users, often referred to as readers, can navigate through it in a more effective and convenient fashion. Additionally, authors can manipulate information according to their needs (Kyriakopoulou et al., 2001). Therefore, hypertext can be regarded as an informal mechanism, which describes the attributes of these objects and captures relationships that possibly exist between them. Such a characteristic made hypertext become known as an alternative way of structuring information.

Autonomous units of data (e.g. text, images, etc.) can be connected non-linearly creating a structure that has the form of a graph. Apparently, such type of organization and representation of information benefits not only the readers but also the authors, each one by their own point of view. More specifically, readers can retrieve the information they want in the right order serving more easily their particular needs, whereas authors can organize their ideas more efficiently by creating relationships (links) between parts of data (nodes). Thus, hypertext can be a precious dialogic means that offers more flexibility and the freedom of choice to the users according to their preferences, the level of comprehension, and other determined factors.

The analogy of a semantic network to hypertext has long been recognized (Conklin, 1987), and a semantic network has been considered as a logical model of hypertext – especially for those hypertexts with typed nodes and links. As it is widely known, a semantic network is a knowledge representation scheme consisting of a directed graph in which conceptual units are represented as nodes, and relations between the units are represented as links. The graph becomes semantic when each node and link is assigned a particular type, making it meaningful. The essential idea of semantic networks is that the graph-theoretic structure of relations can be used for inference as well as understanding (Lehmann, 1992). In this paper we claim that semantic networks may be profitably viewed as hypertext.

Trying to model different user needs in hypertext, the notion of domain appeared, defining special structural abstractions with specific properties as well as a set of behaviors. The role of structural abstractions is to capture and generalize the knowledge of different problem domains, whereas behaviors are described as computation over structure which is considered as a crucial parameter for the semantic of hypertext structure (Leggett and Schnase, 1994) (see table 1). For example, the idea of taxonomic domain was coined by biologists wanting support for the task of creating taxonomies of the species they were researching (Nürnberg et al., 1996). Similarly, within the last decades, various domains, such as navigational (Halasz, 1987), spatial (Marshall et al., 1994), argumentation (Conklin and Begeman, 1987), etc., have emerged. Since semantic networks and hypertext are closely related, the former ones may be considered as a new domain. The issue in hypertext upon the introduction of a new domain is not to express the domain structure using some general model of structure, but to provide users with domain specific structure to directly work with.

Taking the aforementioned into consideration, it is inferred that the need for domain existence in hypertext is essential. Towards the better exploitation of the properties provided by a particular domain, tools can be developed in order to utilize these specific structures. In this way, users can have the opportunity to work with these tools in order to perform syntactic and/or semantic checks, and maybe to perform structural computations that are only relevant within the domain. Therefore, semantic networks can possibly take advantage of these features improving the infor-

Domains	Structural Abstractions	Behaviors
Navigational	node, link, anchor	follow link, generic links
Taxonomic	taxonomy, taxon, specimen	open taxon, compare, auto generate, detect double categorizations
Spatial	item, space, implicit structure	spatial parse
Argumentation	issue, position, evidence	support link, oppose link, circular argument detection
Wordnet	synset	?

Table 1: Example domains in hypertext.

mation management and graph organization.

4. Approaching Wordnet via Hypermedia

Hypertext and wordnets share many common points regarding their fundamental principles and the objectives towards which they both aim. In particular, they are both targeted towards capturing relations that possibly exist between objects and thus providing information of the underlying objects via various types of links used for describing the relations. Therefore, the main characteristic of wordnets and hypertext systems is the ability to create associations between semantically related information items. On the one hand, these associations imply purposeful and important relationships between associated materials, whereas on the other hand the emphasis upon creating associations stimulates and encourages habits of relational thinking of the user (Landow, 1987).

Relations form the notion around which both semantic networks and hypertext are organized. In the case of semantic networks, relations are denoted explicitly between the lexical units they contain via predefined lexical links, and capture information on the semantic properties of words. In the case of hypertext, although the notion of association can be met in all hypertext domains, the navigational domain with the use of *links* is more closely related to it. Consequently, lexical relations form the fundamental entity of semantic networks the same way as associations in hypertext form the basic structural element around which domains are modeled.

In both cases, information objects (either lexical or not) are heavily structured in order to enable users of wordnets or hypertext navigate through the information they contain successfully. Structure is achieved via internal links, which form the basis on which information is stored and expressed. However, links in semantic networks and hypertext are until recently viewed as two distinct elements and no attempt has been made towards comparing the two. We report on the similarities that exist between hypertext relations and semantic links in an attempt to model the latter in hypertext systems.

In order to support this linking activity in an effective way, hypertext researchers have created a flexible link structure incorporating different levels of functionality. More specifically, in hypertext one can create single or bi-directional links, binary or n-ary links, links to links, automatically activated links, etc. Similarly, links in wordnet are bi-directional and there is generally no restriction on the number and types of links they could be included in it as long as the relatedness between the information items is properly and adequately expressed. Bi-directionality of links indicates that if an object A is somehow related to an object B then object B is again related via the same or another relation to the object A.

However, since bi-directionality might not always be the case in wordnets, special tags need to be attached to the relations to denote their single direction. Namely, tags are being used on semantic network relations to indicate that a lexical item is related to another via a particular type of link but not vice versa. Tags are attached to each link separately and act like constraints on the information provided by the link. However, in the case of hypertext, due to the existence of many specialized domains, the notion of tags is used implicitly.

Furthermore, besides creating associations among semantically related information items, another characteristic shared between hypertext and semantic networks is inheritance. This feature implies that properties of the father are inherited to the children. More specifically, the notion of generalization and specialization forms the principle on which relations are expressed. Specialization and generalization define a containment relationship between a higher-level entity set and one or more lower-level entity sets. Specialization is the result of taking a subset of a higher-level entity set to form a lower-level entity set, whereas generalization is the result of taking the union of two or more disjoint (lower-level) entity sets to produce a higher-level entity set.

Inheritance in wordnets is described via the *H/H tree* that is the complementary hypernymy/hyponymy relations. This type of relationship between objects result in viewing wordnets like tree-structured sources of information, and thus not allowing circular loops. As far as hypertext is concerned, these organizational structures exist in the taxonomic domain under the respective terminology of *supertaxon* and *subtaxon*. The subtaxon is associated with the supertaxon via an "is-a" relationship, inheriting all the characteristics that the latter might have. In particular, the user can classify objects (known as specimens) into sets according to their features, search within the members of a set to find relationships or discreet subsets, and create new sets from the already existing ones.

Finally, what should be stressed is that semantic networks and hypertext, despite the characteristics they have in common, they also have quite a few differentiations, mainly stemming from their applications and usage. What we attempted in this paper is to explore the usefulness of both wordnets and hypertext systems beyond the limitations imposed by the applications at which they are targeted. What we claim is that by treating wordnet, as a new domain of hypertext would result in a better understanding of the language structure and consequently human memory and way of thinking. After all, any application is targeted towards human beings and aims at providing a clear description of how information is stored and thus how it should be interpreted. In this respect we strongly believe that if semantic networks are viewed beyond strictly linguistically constraints and applications, the results could only be beneficial.

5. Discussion

As it has been already mentioned, the technology of hypertext is not mainly used for the organization of information but can be considered as a significant means of structuring information. Viewing semantic networks as hypermedia, the power of hypertext is enforced even more, making us infer that any kind of information can be structured under the fundamental characteristics of hypertext. Furthermore, some special structural characteristics of semantic networks can be effectively exploited by hypertext community, resulting in the extension of already existing domains, such as taxonomic, navigational, etc. More specifically, tags might be such a characteristic, providing the hypertext users with the ability to pose semantic constraints upon relations, enabling the distinction among different types of whichever kind links.

On the other hand, taking advantage of the structural characteristics of hypertext while developing semantic networks can prove quite beneficial for both the lexicographic and linguistic communities. In particular, hypertext provides ways of organizing information stored in such systems in a meaningful way so that navigation through the stored data is facilitated. By adopting structures implied by the hypertext community in other applications such as lexicography, the potential and performance of the latter can be greatly improved. When it comes to the storage of lexicographic data the need for efficient structures becomes apparent due to the large amount of information that has to be handled and especially due to the dynamic nature of the underlying information. Moreover, even if behaviors exist in wordnets, they haven't been explicitly defined so far, resulting in less comprehensive usage of the underlying data.

Language forms the mean through which communication is achieved and as such its processing undergoes through various structural decisions that need to be taken prior to storing and incorporating lexicographic data in applications. In this paper we attempted a preliminary comparison among structural characteristics of semantic networks with hypertext and as a conclusion we claim that the abovementioned areas share a few common points in terms of data representation, storage and navigation. What we imply is that semantic networks and hypertext are by no means equivalent in terms of structure. Conversely, what we suggest is that by tracing points between the two and by adopting structural characteristics of other domains can only be beneficial for both sides.

6. References

Jeff Conklin and Michael L. Begeman. 1987. gIBIS: A Hypertext Tool for Team Design Deliberation. In *Proceedings of the ACM Conference on Hypertext*, pages 247–251, Chapel Hill, North Carolina, United States. ACM Press.

Jeff Conklin. 1987. Hypertext: An Introduction and Survey. *IEEE Computer*, 20(9):17–41.

Martha W. Evens, editor. 1988. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University Press, Cambridge, England.

Frank G. Halasz. 1987. Reflections on Notecards: Seven Issues for the Next Generation of Hypermedia Systems. In *Proceedings of the ACM Conference on Hypertext*, pages 345–365, Chapel Hill, North Carolina, United States. ACM Press.

Ruqaiya Hasan. 1984. Coherence and Cohesive Harmony. In James Flood, editor, *Understanding Reading Comprehension*, pages 181–219. IRA.

Martin Kay. 1989. The Concrete Lexicon and the Abstract Dictionary. In *Proceedings of the 5th Annual Conference of the UW Center for the New Oxford English Dictionary*, pages 35–41, Waterloo, Ontario, Canada.

Maria Kyriakopoulou, Dimitris Avramidis, Michalis Vaitis, Manolis Tzagarakis, and Dimitris Christodoulakis. 2001. Broadening Structural Computing Systems Towards Hypermedia Development. In *Proceedings of the 3rd International Workshop on Structural Computing*, pages 131–140, Århus, Denmark. Springer-Verlag.

George P. Landow. 1987. Relationally Encoded Links and the Rhetoric of Hypertext. In *Proceedings of the ACM Conference on Hypertext*, pages 331–343. ACM Press.

John J. Leggett and John L. Schnase. 1994. Viewing Dexter with Open Eyes. *Communications of the ACM*, 37(2):76–86.

Fritz W. Lehmann. 1992. Semantic Networks in Artificial Intelligence. In Fritz W. Lehmann, editor, *Semantic Networks*, pages 1–50. Pergamon Press Ltd.

Catherine C. Marshall, Frank M. Shipman, and James H. Coombs. 1994. VIKI: Spatial Hypertext Supporting Emergent Structure. In *Proceedings of the 1994 ACM European Conference on Hypermedia Technology*, pages 13–23, Edinburgh, Scotland. ACM Press.

Catherine C. Marshall. 1987. Exploring Representation Problems Using Hypertext. In *Proceedings of the ACM Conference on Hypertext*, pages 253–268, Chapel Hill, North Carolina, United States. ACM Press.

George A. Miller. 1998. Nouns in Wordnet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 23–46. MIT Press.

Peter J. Nürnberg, John J. Leggett, Erich R. Schneider, and John L. Schnase. 1996. Hypermedia Operating Systems: A New Paradigm for Computing. In *Proceedings of the the 7th ACM Conference on Hypertext*, pages 194–202, Bethesda, Maryland, United States. ACM Press.

Peter J. Nürnberg, John J. Leggett, and Erich R. Schneider. 1997. As We Should Have Thought. In *Proceedings of* *the 8th ACM Conference on Hypertext*, pages 96–101, Southampton, United Kingdom. ACM Press.

Uta Priss. 1998. The Formalization of Wordnet by Methods of Relational Concept Analysis. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 179–196. MIT Press.

Requirements for Domain-Specific WordNets

Koutsoubos Ioannis-Dimitrios, Christodoulakis Dimitris

Computer Engineering and Informatics Department,
University of Patras
Research Academic Computer Technology Institute
koutsoub@ceid.upatras.gr
dxri@cti.gr

Abstract

This paper addresses the need for domain-specific resources in NLP applications. The motivation for this work emerged from the current limitations of WordNet when the latter is adopted in a domain-specific applications and environments. Moreover, we report on methods and techniques for extending and tuning WordNets for domain-specific usage. We envisage a unifying WordNet structure, that will be easily extendable and customizable and also has the ability of incorporating other lexical and semantic resources with minimum effort. Finally, we discuss on the advantages of a unified WordNets structure over various types of applications that require extensive usage of NLP applications.

1. Introduction

Lexical resources used in natural language processing have evolved from handcrafted lexical entries to machine-readable lexical databases and large corpora. Much effort is being applied no the creation of electronic lexicons and electronic linguistic resources in general. However, the above resources are expensive to build, and instead of creating new ones from scratch, it is preferable to adjust and extend existing ones.

One linguistic resource of great interest is WordNet (FellBaum, 1998). WordNet is a general-purpose concept ontology, which has been developed a Princeton University, and resembles the way that humans store and organize information in their memory. It can be used both as an on-line dictionary or thesaurus for reference purposes, and as a taxonomic lexical database. WordNet is a resource of high quality and is freely available over the Internet, thus it has rapidly become one of the most widely used tools in language engineering, research and development.

However, as many technical words or word meanings cannot be found in general semantic databases such as WordNet, Natural Language Processing (NLP) in specific domains requires specialized semantic lexica. A major difficulty in using WordNet or any other general NLP resource in a specific domain is that much of the specialized semantic attributes (terminology, semantic relations, domain-specific relations etc) of the domain is not present. In this paper we describe the requirements of domain-specific wordnet development. First, we describe shortly the application usage of wordnet. We then explain the need for domain-specific NLP resources. Next we present techniques and methodologies that are used till now for the development of domain-specific wordnets. Finally, we present our approach towards developing domain-specific wordnets. Finally, we outline some early conclusions regarding the necessity for building domain specific WordNets and their usefulness in various applications.

2. WordNet applications

WordNet has been identified as an important resource in the human language technology and knowledge processing communities. Its applicability has been cited in many papers and systems have been implemented using WordNet. Almost every NLP application nowadays requires a certain level of semantic analysis. The most important part of this process is semantic tagging: the annotation of each content word with a semantic category. WordNet gives a solution to the above problem and has been used in various applications including Information Retrieval, Word Sense Disambiguation, Machine Translation, Conceptual Indexing, Text and Document Classification and many others.

3. Need for domain-specific resources

A problematic issue is that general semantic resources like WordNet do not cover many terms and concepts specific to certain domains, and also include many unnecessary (general) concepts and relations. Therefore these resources need to be tuned to a specific domain at hand. This involves selecting those senses that are most appropriate for the domain, as well as extending the sense inventory with novel terms and novel senses that are specific to the domain (Buitealar, 2001; Turcato et al.,2000). Another problem is that in a specific domain only a subset of the semantic relations defined in the general semantic resource hold. Also many technical words or word meanings cannot be found in general resources. Partial overlaps can be found, but the domain specific description is likely to be more precisely defined and reliable.

As a result of these difficulties with existing generic resources, NLP system builders have tended to handcraft resources for each application domain, or have looked at techniques for automatically or semi-automatically constructing lexicons of various sorts from texts in the domain.

The main problem is how can we develop domain-specific resources either from scratch or by using existing resources with minimum effort.

There are two main problems. The extension/expansion of existing general resources and the adaptation of these resources to a specific domain and how can we acquire the above with minimum effort. In particular, the first problem regarding extending already existing lexicographic resources with domain-specific terminology requires a lot of manual work since additional information needs to be attached to the contents of the resources emerging from the underlying domain of interest. This would imply that large corpora from various terminological domains should be used in order to perform a semantic annotation of the terms they comprise of. In the second case, adapting existing resources toi particular applications would require not only enriching those resources with specialized terminology but it would also need partial restructuring of the resource so that the new content is sufficiently represented in a meaningful way.

In the case of wordnets the solution that implies the development of domain specific semantic networks seems as the best way of solving many problems imposed by the lack of such resources from various NLP applications. In the following sections we briefly report on the work conducted so far in this area and we continue with a description of our approach towards the necessity of domain specific terminological resources.

4. Building Domain-Aware WordNets so far

It is obvious from the above how important is the need for domain-specific NLP resources in general, and particularly for domain-specific wordnets. Several methods for the creation of domain-specific resources have been applied ranging from:

- Creation from scratch, to
- Data Extension of generic WordNets for a specific domain and
- Structure Extension of generic WordNets for a specific domain.

More specifically, the methodology adopted for each of the abovementioned techniques is described as follows:

4.1. Creation from Scratch

One solution, and apparently the most costly, is to handcraft domain-specific wordnets from scratch for any specific-domain. Building wordnets by hand requires significant amount of time and effort even for restricted domains. Furthermore this effort is repeated when a system is ported to another domain. The above leads us to automatic or semi-automatic approaches for building wordnets and other NLP resources using already available existing generic resources.

4.2. Data Extension of generic WordNets for a specific domain

The adaptation of existing resources to a specific domains includes selecting those terms and meanings that are relevant for the domain, adding new terms and meanings that are missing from the existing resource, removing relations that are irrelevant or incorrect in the specific domain, keeping relevant relations and adding missing ones (Buitelaar & Sacaleanu, 2001,2002;Turcato et al.,2000).

4.3. Structure Extension of generic WordNets for a specific domain

Another solution to the problem is to extend existing generic wordnet structure incorporating in it semantic distinctions from external resources such as ontologies, semantic taxonomies, domain-specific corpora etc. One approach is to add an ontology layer, which refers to specific domain attributes and characteristics and thus relates the domain with the linked concepts (Vossen, 1998; O'Sullivan et al., 1995). Another way is to link concepts with relevant document collections or corpora and find a way to compute the weights of their topic signatures (Agirre et al., 2001). Automatically build an hierarchy of terms using terms extracted from documents of a specific domain, combine it with existing hierarchies in wordnet and by fusing and clustering we can derive a condensed tree that has maximum coverage due to the extension, but only contains distinctions and classifications that are relevant and desired (Vossen, 2001).

There have also been attempts to integrate the information of generic lexical databases with existing ones (Magnini & Speranza, 2001).

5. What is missing from WordNet?

The success of WordNet has determined the emergence of several projects that aim the construction of WordNets for other languages than English or to develop multilingual or specialized WordNets or to extend existing WordNets for specific domains or to incorporate WordNet in various NLP applications. Through these attempts many WordNet's advantages have been discovered and some weaknesses have appeared. According to *(Harabagiu et Al. 1999)* the main weaknesses of WordNet cited in the literature are:

1. The lack of connections between noun and verb hierarchies.
2. Limited number of connections between topically related words.
3. The lack of morphological relations.
4. The absence of thematic relations/ selectional restrictions.
5. Some concepts (word senses) are missing.
6. Since glosses were written manually, sometimes there is a lack of uniformity and consistency in the definitions.

Until now there has been a lot of research for methods and techniques for WordNet development, customization, multilinguality, alignment with existing resources etc. However all the attempts concentrated on everything that was related to the content of WordNet and WordNet's lexical and semantic coverage, leaving behind everything that is related to the data model of Wordnet and WordNet's structure (the way that WordNet's data are stored and manipulated).

From a WordNet's developer perspective the main disadvantage of WordNet is that WordNet is almost a black box. The WordNet community is increasing year by year, but till now there are no standards about WordNet structure. With a standard WordNet structure and all the methods and techniques that are already available for WordNet construction, extension, alignment with other NLP resources and link with other language WordNets will road the map for a new perspective towards wordnets and their usage in every day NLP applications.

6. Requirements for Domain-Specific WordNets

In this section we describe the requirements that a domain-specific WordNet must satisfy. Many of these requirements are also addressed to generic WordNets.

One key point is the integration of domain-specific wordnets with generic ones. On the one hand the domain-specific wordnet is a specialized resource, whose content is supposed to be more accurate and precise for the domain that it was designed; on the other hand, the generic wordnet guarantees a more uniform coverage as far as high level senses are concerned. There must be a flexible and modular integration procedure, which will give the ability many domain-specific wordnets to co-exist with one generic one. This procedure shall manage inconsistencies and overlaps between the different resources. Co-existence of lexical resources that are targeted towards various domains has many advantages.

First and foremost, it enables the comparison of concepts used in both genetic and domain specific vocabulary. ,It can also contribute towards the ease identification of the domain in which a concepts belongs to. However, the most important feature of such resources is the potential of using a domain specific semantic resource for various types of applications. The latter results in a global lexicographic resource of great usefulness in many tasks and applications.

A problematic issue in the field of NLP is that it does not often suffice to depend on any single resource , either because it does not contain all required information or the information is not organized in a way suitable for the purpose. So merging of different resources is necessary. Many different NLP resources are available to the NLP community e.g. corpora, morphological lexicons, semantic lexicons, ontologies. Many applications will benefit from the integration of such resources with WordNet (Kwong, 1998). So there shall be a flexible structure that will provide fully-automatic or semi-automatic mechanisms for the incorporation of such resources in WordNet.

WordNet has been criticized for its lack of relations between topically related concepts. The enrichment of WordNet's concepts with topic signatures and the application of topic relations open the avenue for interesting ontology enhancemenrs, as they provide concepts with rich topical information (Agirre et al., 2001). For instance, similarity between topic signatures could be used to cluster topically related word meanings. Word sense disambiguation methods could profit from these richer ontologies, and improve word sense disambiguation performance.

WordNet's concepts shall be enriched with additional semantic and non-semantic attributes. Some of these attributes may be word usage examples, words that accompany a concept in a specific meaning, morphology information, domain-specific information about the concept. For example if we meet the word 'world' with the meaning of 'people' we cannot find this word in plural. The above attributes may also be links to corpora or other incorporated resources. By the same way attributes shall be applied to relations, too. For instance some relations may exist under certain constraints in a domain-specific context, and there must be a way of identifying domain-specific relations that do not exist in generic or in other domain contexts, or generic relations that are also applied in domain-specific context. One such examples concerns the application of wordnet during language teaching tasks in which phonetic information could be added.

The WordNet structure shall be organized in a way that will allow the insertion of additional relations between concepts, additional attributes both for concepts and relations and constraints both for attributes and relations without affecting existing data and with a way that will be as easy and effective as possible.

Another feature that shall be made available to WordNet is the definition of the behavior of relations regarding the domain that the wordnet is designed for and the application usage of the WordNet. Following this approach different applications in a specific domain have the ability to share common data. This means that if somebody developed a domain-specific WordNet for domain A in order to use it in his document classification application and another one plans to develop a query expansion system for an information retrieval application he can use the already developed WordNet in the same way only by changing for instance the behavior of the synonymy relation which will now be used for searching in documents with the synonyms of a given word. In other applications for instance the hyperonym relation may be used for getting more general word meanings than the given one and in other applications may define an upper-level category of classification of documents.

An Ontology Layer should be present on the upper level of the semantic features of a language for the transfer of domain specific semantic characteristics and distinctions relative to the domain to the underlying concepts. However, it might be more effective if the concepts belonging to the upper level had as additional features the abovementioned distinctions and thus all terms related to these inherit these distinctions and features. . The above resembles much the wordnet-type of information storage and representation and would result in a more flexible semi-automatic extraction and development of domain-specific ontologies based on wordnet information.

All the above leads us to the conclusion that there is an imperative need for a flexible and unifying WordNet structure. The whole WordNet community shall concentrate in the standardization process of WordNet structure. The structure must be able of defining concepts, relations, attributes for both of them, flexible linking with existing NLP resources and components. It also must be easily customizable and extendable, allow the co-existence of generic and specialized wordnets providing mechanisms for domain resolution and identification. Such an approach will make easier the process of multilingual wordnet linking and will also provide an unifying approach to any NLP problem that wordnet is called to solve. Since the research concerning wordnet itself and its applications has grown extremely in the past years a standard structure will just provide wordnet an easy and effective way in everything concerning wordnet from wordnet development to wordnet usage in NLP applications.

With the need of the standardization of structure comes the need for a wordnet protocol, which will describe all the operations, methods, functions that wordnet offers. The existence of a wordnet protocol means that everyone is free to develop wordnet in the way they prefer even if it is relational databases, xml files, polaris format files, indexed text files etc, as long as they follow the pre-specified protocol.

The need for a unified structure is requested to solve problems related to wordnet extension as long as other problems emerging from wordnet applications ands need to be solved via a unified and common way. The main idea behind this assumption is the conversion of wordnets into a linguistic resource that would apply to as much as possible to ll members of the NLP community.

The need for a common protocol needs to be solved through a unification of the applications of the already existing wordnets. A common protocol applications envisaged for one monolingual wordnet (e.g. the English Wordnets) could be used in other monolingual wordnets without any previous change required in the structure or content of the latter. Of course this implies that in case one application performs sufficiently for a particular domain then its usage in another domain needs solely the existence of a wordnet for another domain and no extra effort towards structural or content modifocations.

7. Conclusion

We identified the need for domain-specific WordNets and presented some requirements that shall be met both by generic and specialized WordNets. WordNets success in the field of NLP can be even greater but to achieve this there must be standardization concerning both the structure and the protocol, which will be used by applications that use WordNet. This is a long way, and it must be walked with the right steps.

8. References

Agirre E., Ansa O., Martinez D., Hovy E.(2001). Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL worshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. Pittsburg, USA

Buitelaar P., Sacaleanu B.(2002). Extending Synsets with Medical Terms In : *Proceedings of the First International WordNet Conference*, Mysore, India.

Buitelaar P., Sacaleanu B. (2001). Ranking and Selecting Synsets by Domain Relevance In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop*, Carnegie Mellon University, Pittsburgh.

Farreres, G. Rigau, and H. Rodriguez (1998). Using WordNet for Building WordNets. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montr'eal, Canada.

FellBaum Christiane (1998). WordNet: An Electronic Lexical Database. MIT Press Books.

Habert B., Nazarenko A., Zweigenbaum P., and Bouaud J.. (1998). Extending an Existing Specialized Semantic Lexicon. In *Proceedings of first International Conference on Language Resources and Evaluation*, pages 663--668, Granada.

Harabagiu S.M.,Miller A. G. and Moldovan (1999). WordNet 2 – a Morphologically and Semantically Enhanced Resource. In *Proceedings of SIGLEX-99* (pp. 1--8). University of Maryland.

Kwong, Oi Yee (1998). "Aligning WordNet with Additional Lexical Resources". In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. Montreal, Canada, August.

Magnini, Bernardo and Manuela Speranza (2001). Integrating Generic and Specialized Wordnets. In *Proceedings of the Conference on Recent Advances in Natural Language Processing, RANLP 2001*, Tzigov Chark, Bulgaria.

O'Sullivan D., A. McElligott, R. Sutcliffe (1995). Augmenting the Princeton WordNet with a Domain Specific Ontology, in *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95).* Montreal, Canada.

Turcato D., Popowich F., Toole J., Fass D., Nicholson D. and Tisher G. (2000). Adapting a synonym database to specific domains. In *Proceedings of the ACL'2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval*. Hong Kong.

Vossen P. (2001) Extending, Trimming and Fusing WordNet for Technical Documents. In: *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop*, Carnegie Mellon University, Pittsburgh.

Vossen, P (ed.) (1998) EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document

Notes about Labelling Semantic Relations in Estonian WordNet

Kadri Vider

University of Tartu,
Department of General Linguistics
Tiigi 78-204, 50410 Tartu, Estonia
kvider@psych.ut.ee

Abstract

Estonian language is rich in derivation. Most of derivational suffixes have their regular meaning(s) and is very obvious, that source and target words in derivation have regular lexical-semantic relations between them. The problem of what regular derivational suffixes in Estonian lexica cover what kind of semantic relations in Estonian WordNet is discussed in this paper.

Another problem of labelling connected with semantic relations is related to proper nouns. In purpose to use referential character of proper nouns in word sense disambiguation, we need to connect proper nouns with objects carrying the names e.g. 'John ISA man', but not 'John ISA first name'.

1. Introduction

Compilation of Estonian WordNet (EstWN) started in 1997 and the work is still in progress. The work was funded partly by the Estonian Science Foundation and partly in the framework of the Estonian National Programme of Language Technology. Like other wordnets, EstWN is a lexical-semantic database, the basic unit of which is concept. Concepts are represented as synonym sets (synsets) that are linked to each other by semantic relations. In 1998-1999 EstWN was created as a part of EuroWordNet (EWN) and since then we have used semantic relations from EWN, which are more flexible and richer than in the original (Princeton) WordNet. Still, our experience has shown that there are at least some language-specific semantic relations needed. Up to now, the usage of semantic relations was limited by the set provided by Polaris, the EWN editing tool.

Which new words or concepts should be concentrated on to upgrade the EstWN? It is essential that words actually used in text will be added. Results of word sense disambiguation (WSD) task of corpus texts turned out to be a good way of adding missing and new synsets and senses into our wordnet. (Kahusk and Vider, 2002)

Estonian is usually considered to be an agglutinative language, thus belonging to same group as Finnish, Hungarian and Turkish. It is flective language with free word order.

In order to reach the lemma in the text, Estonian needs morphological analysis. The program ESTMORF, in use at present, renders it possible to analyse the productive derivatives and tag suffixes.

2. Semantic relations in EstWN

The existing Estonian WordNet contains nouns, verbs, some adjectives and proper nouns, more than 10,000 synsets all together. The more detailed description of EstWN is given in the final document of EuroWordNet, Estonian part (Vider et al., 1999)

Semantic relation	**No. of links**
has_hyperonym/has_hyponym	19002
belongs_to_class/has_instance	948
near_synonym	354
xpos_near_synonym	246
has_holonym/has_meronym	234
antonym	209
be_in_state/state_of	186
near_antonym	138
involved/role	134
causes/is_caused_by	128
has_subevent/is_subevent_of	36
has_xpos_hyperonym/has_xpos_hyponym	12
xpos_near_antonym	4
xpos_fuzzynym	2

Table 1: Semantic relations expressed in EstWN in the order of frequency.

3. System of Estonian derivation

Wordnet is based on word meaning and from this point of view such lexical feature as derivation should not play a significant role. But a lot of Estonian derivational suffixes have concrete meanings and this fact can be applied in connecting the derivational base and the derivation with a definite semantic relation, dependent on the meaning of the derivational affix.

In Estonian, derivation is mainly a process of appending derivational suffixes, more than 60 altogether, to both declinable and conjugable words. Suffixes can be appended sequentially; up to four suffixes in a row can be appended in some cases. About 8% of the word forms in a running Estonian text are derived words; in journalism and scientific texts the figure is even higher (Kaalep, 1997).

Derivation, a frequent and productive way in Estonian for forming new words, is a process where adding an affix produces a new lexical item having its own inflectional paradigm. Derivational morphology in Estonian is always connected with changing the meaning of lexeme. The lexical meaning of the derived word is different from the word used as the derivational base, in some productive cases the derived words belong to a different part of speech

Thus it may be concluded that affixes in Estonian belong to the category of semantics, not grammar. Morphologically derivation can be defined as the formation of a new stem by adding an affix to the last morpheme of the stem.

In Estonian, compounding is even more frequently used for word formation than derivation. Compound words comprise more than 12% of the running words in an average Estonian text. The formation of Estonian compounds is quite free and derived words may also constitute a compound. In this paper we consider only such kind of compounds.

We proceed from the assumption that in a lexicon compiled on the semantic basis the semantic association between the words derived from the same stem should be fixed. It should be possible to automate the relation on the basis of meaningful affixes. The question is which relation should be attributed to which affix.

Derived/derived_from/has_derived relations exist in EWN structure (Vossen 1999), but they are clearly too general and ambiguous for such a language abundant in regular and ample derivation as Estonian.

4. Suffixes actual in EstWN data

This chapter deals, first and foremost with the productive derivation types (formation patterns) that have an independent meaning, e.g.

VERB+**mine** – PROCESS[NOUN],
VERB+**ja** – ACTOR,
PLACE[NOUN]+**lane** – INHABITANT

Lexicalised derivation also has quite a clear relation with the derivational base. Only the idiomised derivations have lost the distinct relation with the derivational base (Kasik 1996).

4.1. Verb suffixes

4.1.1. Verb -> Verb derivation

Most frequent verb suffix in Estonian is -**ta**, which has a causative meaning in verb-to-verb derivation, e.g. kulu/ta/ma (spend, expend) *causes* kuluma (go, be spent); levi/ta/ma (distribute, cause to spread) *causes* levima (spread, be disseminated); liigu/ta/ma (cause to move) *causes* liikuma (move); kao/ta/ma (lose, fail to keep) *causes* kaduma (disappear, vanish, get lost); meenu/ta/ma (remember, retrieve, recall, remind) *causes* meenuma (be reminded); kuiva/ta/ma (dry, make dry) *causes* kuivama (become dry); sünni/ta/ma (birth, give birth) *causes* sündima (be born); nõrges/ta/ma (weaken, make weak) *causes* nõrge/ne/ma (weaken, get weak); aren/da/ma (develop, evolve) *causes* are/ne/ma (evolve, undergo an evolution); puru/sta/ma (break, cause to break) *causes* puru/ne/ma (break, separate, be smashed); rahu/sta/ma (calm, make calm) *causes* rahu/ne/ma (calm, be pacified, become stable); unu/sta/ma (forget, fail to remember) *causes* unu/ne/ma (pass out of mind, be forgotten).

Productive verb suffix **–u** constructs intransitive verbs with reflexive meaning, eg. aeglus/ta/ma (retard) - aeglust/u/ma (slow, become retarded); asen/da/ma (substitute, replace) - asend/u/ma (be replaced); eral/da/ma (separate, divide) - erald/u/ma (separate from); eru/ta/ma (stimulate, shake, excite) - erut/u/ma (become excited about); eten/da/ma (perform, give a performance) - etend/u/ma (play, be performed); kahjus/ta/ma (damage, do harm) - kahjust/u/ma (be damaged); katma (cover) – katt/u/ma (be covered); kuhjama (heap, pile, stack) - kuhj/u/ma (be heaped, be piled); moodus/ta/ma (form, constitute) – moodust/u/ma (be formed, be constituted).

The most important derivation in this group is muutma (*change, alter, make different*) - muut/u/ma (*undergo a change, become different*). The source verb of derivation needs an active agent, but it does not render passive or *is_caused_by* meaning to the verbs with reflexive u-suffix. Lexical expression of passivity is not characteristic of the Estonian language. As to Estonian (perhaps French and German as well) reflexivity is one of the missing semantic relations in the EWN verb structure.

4.1.2. Noun -> Verb derivation

The most common semantic categories in derivations of this type are CAUSE, CHANGE, USE, ADD. Verb arguments behave in this case as derivatives, e.g. RESULT, ACTOR, INSTRUMENT. They all hold subtypes of *involved/role* relation. Often such arguments can be met in synonymous phrases or idioms, e.g. kirju/ta/ma, kirja panema (*write, write down, directly "put into letter"*).

(1) Productive suffix **-ta** and its variant -**sta** have factitive meaning, i.e. one of the arguments of the derived verb is the source of derivation as well. The semantic relation between the verb and its derivational base belongs, in this case, to the subtype of *involved/role* relation, e.g. huvi/ta/ma (interest, cause to be interested) *involved* huvi (interest); avar/da/ma (enlarge, expand, extend) *involved* avar (spacious); elav/da/ma (enliven, liven) *involved* elav (living, alive); nalja/ta/ma (joke, jest) *involved* nali (wit, humour, joke, jest); ahel/da/ma (chain) *involved* ahel (chains, chains); halven/da/ma (make worse, worsen) *involved* halb (bad)

(2) Suffixes **–u** and **–ne** have translative meaning. They present autonomic CHANGE (of state or situation); e.g. korts (wrinkle, fold, crease) – korts/u/ma (wrinkle, ruckle, crease, crinkle, scrunch) kõva (hard, firm, solid, stiff) – kõvast/u/ma (harden, indurate, solidify); kõver (crooked, bent, curved) - kõverd/u/ma (curve, crook, bend); külm (cold) – külm/u/ma (freeze, change to ice); lahus (solution) – lahust/u/ma (dissolv, resolve); niiske (damp, moist) – niisk/u/ma (moisten, dampen); puit (wood) – puit/u/ma (turn into wood, lignify); raev (rage, fury)- raev/u/ma (become furious, see red); rasv (fat, lardy) - rasv/u/ma (fatten, batten, grow fat); rohi (grass) - roht/u/ma (overgrow with grass); suund (direction) – suund/u/ma (head, travel in a direction); kitsas (narrow) – kitse/ne/ma (narrow, contract); halb (bad) – halve/ne/ma (worsen, decline); harv (sparse, thin) – harve/ne/ma (thin out); kauge (far) – kauge/ne/ma (recede, move away)

Existential verbs, where derivation changes only the part of speech should be brought out as a separate group.

4.1.3. Modifying derivation

Derivations formed with the help of affixes modifying the verb have a hyperonym/hyponym relation with the derivational base, for the affixes mentioned above only modify the way of action. The best label for describing such a relationship is troponymy.

Frequentatives (expressing repetition of an action, e.g. hüppama (jump)- hüp/le/ma (hop, skip, jump lightly); mulks (gurgle) - mulks/u/ma (bubble up); tukse (throbbing) - tuks/u/ma (pulsate, throb, pulse); momentanes (express the singleness or suddenness of an action, e.g. tuks/u/ma (pulsate, throb, pulse) - tuks/ata/ma (give a throb)) and continuatives (show the continuity and permanence of an action, e.g. mängima (play) - mängi/tse/ma (dally, trifle, play)) can be differentiated by the affixes.

4.2. Noun suffixes

In case of argument-nominalization the derivative is expressed in the function of one argument of the derivational verb. The more widely-spread arguments include ACTOR, RESULT, INSTRUMENT, OBJECT, PLACE.

4.2.1. Action derivatives

The suffix of absolute productivity **-mine** changes only the part of speech of the derivational base. With the help of this suffix every verb can be changed into a noun, which has cross-part of speech synonym relations, e.g. alustama (begin, start, commence) *xpos_near_synonym* alusta/mine (beginning, start, commencement); harjutama (drill, exercise, practice) *xpos_near_synonym* harjuta/mine (practice session, exercise).

Abstract and metaphorical meanings of the verb should not be bound to the suffix -**mine** but only the ones expressing a definite action.

Due to absolute productivity we have included only such mine-derivatives in the EstWN that were founded in corpus texts.

4.2.2. Personal derivatives

Actor's suffix **–ja** is also a very productive suffix, the application of which is universal for all kind of action, e.g. ehitama (build, construct, make) *involved_agent* ehita/ja (builder, constructor); esindama (represent, be a delegate for) *involved_agent* esinda/ja (representative); juhatama (head, lead) *involved_agent* juhata/ja (leader); kasvatama *involved_agent* kasvata/ja, koristama *involved_agent* korista/ja, kütma *involved_agent* küt/ja, laulma *involved_agent* laul/ja. Some of the ja-derivatives can besides the live agent also express appliances, e.g. ajamõõt/ja (timekeeper); voolumõõt/ja (ammeter); raadiosaat/ja (radio transmitter).

The most productive affix in forming generic names from proper names is **-lane**. The biggest group of lane-derivatives refers to persons by their origin, e.g eest/lane (Estonian); ameerik/lane (American); hiin/lane (Chinese); indiaan/lane (American Indian).

Terms of biological taxonomy form another big group, which could be formed with the help of suffixes **-lane** e.g. kass (cat) - kas/lane (feline, felid); koer (dog) - koer/lane (canine, canid); and **-line**, e.g. kabja/line (perissodactyl mammal); kiletiiva/line (hymenopterous insect); kõrre/line (graminaceous plant).

A productive affix in forming business titles is **-ur**, e.g. kaevama (dig) - kaev/ur (digger, miner); kala (fish) - kal/ur (fisher, fisherman); juus (hair) - juuks/ur (hairdresser); valvama (protect) - valv/ur (defender, guardian, protector).

Feminine suffixes **-nna, -tar** are also productive, e.g. luuleta/ja (poet) - luuleta/ja/nna (poetess); sõber (friend) - sõbra/nna, sõbra/tar (girlfriend). Estonian morphology lacks feminine markers, feminine suffixes exist only in noun derivation. The problem is not new, as in his first papers about EWN-1 Vossen declared that the semantic category WOMAN got lost in converting the Vlis (Dutch) database relations into EWN ones.

4.2.3. Place and set derivatives

All **-la** derivatives refer to a place and indicate a specific place (building, room), e.g. haige (sick person, sufferer, patient) *involved_location* haig/la (hospital); levima (spread, be disseminated) *involved_location* levi/la, parkima (park) *involved_location* park/la (parking lot, car park); suvitama (summer) *involved_location* suvi/la (summer house); sööma (eat, take in) *involved_location* söök/la (lunchroom, eating house).

-kond is a productive suffix expressing collectivism, e.g elanik (inhabitant) *has_holo_member* elanik/kond (population); inimene (human, man) *has_holo_member* inim/kond (humankind, mankind); võistleja (contestant) *has_holo_member* võist/kond (team, squad).

Apart from the kond-suffix, suffix **-stik** refers to the group or set of things or fenomena, e.g. kõrge (high) *has_holo_member* kõrgu/stik (highland, upland); leht (leaf) *has_holo_member* lehe/stik (leafage); mägi (mountain, hill) *has_holo_member* mäe/stik (mountain range); nimi (name) *has_holo_member* nime/stik (list, listing); rahvas (people) *has_holo_member* rahva/stik (population); seade (mechanism) *has_holo_member* seadme/stik (machinery, equipment); taim (plant, plant life) *has_holo_member* taime/stik (vegetation, flora).

4.2.4. Property derivatives

Productive suffix **-us** makes it possible to form property names from most of the adjectives, changing only the part of speech, e.g. intensiiv/ne (intense) – intensiivs/us (intensity, intensiveness); musikaal/ne (musical) – musikaals/us (musicality, musicalness); soola/ne (salty, salt) – soolas/us (saltiness, salt); keeru/line (baffling, knotty, problematic) – keerulis/us (complexity, complexness); lopsakas (buxom, chubby, plump) – lopsak/us (fleshiness, obesity); vürtsikas (hot, spicy) – vürtsik/us (spicery, spiciness)

Suffix **-ndus** forms abstract names of substances or fields of action from concrete nouns, e.g. kauba/ndus (commerce); kirja/ndus (literature); koka/ndus (cookery, cooking, cuisine); maja/ndus (economy); metsa/ndus (forestry); teeni/ndus (service); veondus (transportation, shipping).

4.3. Adjective suffixes

It is difficult to group adjective suffixes by meaning because most of the suffixes can express several meanings. Very often it is dependent on the derivative base.

The adjectives formed from the nouns often convey a comparative or possessive meaning, e.g analoogia (analogy) – analoogi/line (analogous); kriitika (criticism, critique) – kriiti/line (critical); värv (colour) – värvi/line (coloured); kasu (use, good) – kasu/lik (useful); noorus (youth) – noorus/lik (youthful).

The EWN derivational relations *derived/has_derived/derived_from* and *pertains_to/is_pertained_to* are namely prescribed for adjective suffixes.

5. Semantic relations of proper nouns

The main inspiration for our WSD system *semyhe* is Agirre and Rigau (1996) similar system that disambiguates the English noun senses based on WordNet hyponym/hypernym hierarchy, taking into consideration the distances between the nodes corresponding to the word senses in the WordNet tree as well as the density of the tree (Vider and Kaljurand, 2001).

In order to improve the operation of the program, the density of the words, that will be disambiguated should be increased. Up to now proper nouns were left out of disambiguation and they comprised 30% of the 0-analysed nouns. As our WSD system uses EstWN, it is essential that proper names encountered in the texts be added to it. Fortunately the EWN database structure includes a type of entry meant for proper names — *word_instance*.

Hyponymy is a relation between classes of entities. Individual entities, presented in texts as proper nouns and in EWN structure as *word_instance* entries, can also be said to belong to some class. To distinguish this relation from hyponymy it is labelled *has_instance/belongs_to_class* in EWN (Vossen, 1999). It is good because it makes it also possible for the WSD system to find out referee among *word_meaning* entries. Thus WSD system can make more precise decisions about the right word meaning, because meaningful context is more dense. Therefore we added all proper nouns existent in the WSD training corpus to EstWN and linked *belongs_to_class/ has_instance* relation to *word_meaning* entries (see Table 1).

Now the question is which proper noun links to which *word_meaning* entry. It seems only natural to link e.g. capital *has_instance* Tallinn, river *has_instance* Volga. It is also possible to link e.g. male, male person *has_instance* John. But is it right to link family *has_instance* Smith, for family refers to a social group, not a person?

Most proper nouns listed in the EstWN refer to a person. The next group as to the frequency is toponyms that refer to a location or place (city, state, land, region) or natural objects (river, mountain, lake, island etc).

6. Conclusions

In the Estonian language derivation is not a feature of morphology. As to the richness of meaning of the Estonian derivation system, the semantic relations existent in the EWN and labeled as *derived/has_derived/derived_from* clearly too scarce. Making use of the recognizability of the suffixes, it is possible to link the derived words with the derivational base words (semi)automatically, specifying the semantic relation on the basis of the meaning of the derivational suffix.

Specifying the semantic relation of proper nouns is of vital importance to increase the conceptual density in solving the wordnet-based WSD task. One should only be careful and persistent in achieving the target concept.

References

Agirre, E. and Rigau, G. 1996. Word Sense Disambiguation using Conceptual Density. In *COLING-96*

Kaalep, H.-J. 1997. An Estonian Morphological Analyser ant the Impact of a Corpus on Its Development. *Computers and the Humanities*, 31:115-133.

Kahusk, N. and Vider, K. 2002. Estonian Wordnet benefits from word sense disambiguation. In *Proceedings of the First International Global Wordnet Conference* (pp. 26-31). Central Institute of Indian Languages, Mysore, India.

Kasik, R. 1996. Eesti keele sõnatuletus. Tartu Ülikooli Kirjastus.

Vider, K., Paldre, L., Orav, H. and Õim, H. 1999. The Estonian Wordnet. In C. Kunze, editor, *Final Wordnets for German, French, Estonian and Czech*. EuroWordNet (LE-8328), Deliverable 2D014.

Vider, K. and Kaljurand, K. 2001. Automatic WSD: Does it Make Sense of Estonian? In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems* (pp. 159-162).

Vossen, P. (ed). 1999. EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document. http://www.hum.uva.nl/~ewn/docs.htm

RussNet: Building a Lexical Database for the Russian Language

Irina Azarova, Olga Mitrofanova, Anna Sinopalnikova, Maria Yavorskaya, Ilya Oparin

Applied Linguistics Department, Philological Faculty, Saint-Petersburg University
Universitetskaya nab. 11, Saint-Petersburg, Russia
azic@bsr.spb.ru , asinopalnikova@yahoo.com, yav_mas@hotmail.com

Absract

The paper describes the on-going work on creating the WordNet-type lexicon for Russian, so called RussNet. The project started 3 years ago, preliminary results will be available at www.phil.pu.ru. The existing database contains verbs, nouns, and adjectives, the number of senses amounting to 2500.

The Top Ontology of RussNet is under construction, it will be co-ordinated with that of EuroWN. RussNet has inherited EuroWN language-internal relations. Several types of derivational links are added to describe Cross-Part-Of-Speech relations as well as Inner-Part-Of-Speech ones. Adjective-to-noun and verb-to-noun relations of words in collocations are described in details.

An overview of methods used for construction of the Russian WordNet is presented, the procedure of sense definition generation is also discussed.

1. RussNet Structure

1.1.Vocabulary

For the RussNet structure we accepted the general approach, presenting only **Standard Russian** lexis, as opposed to various terminological subsets. The position doesn't prevent us from including those terms that were incorporated into the common language.

On the one hand this approach follows Russian lexicography tradition and on the other hand allows us to provide first and foremost **frequently-used current** vocabulary, that will be exploited by the majority of users. The main sources for such words are newspaper and magazine articles.

1.2. Inherited Features in RussNet

- RussNet is structured along the same lines as Princeton WN, EWN (Vossen, 1998, Miller et al, 1993) and other wordnets: words are grouped into synonym sets (**synsets**), each representing one underlying concept.
- Synsets in their turn are linked by means of various **Language Internal Relations** (LIR), such as hyponymy/hyperonymy, antonymy, meronymy/holonymy, entailment, causation, etc., hyponymy/hyperonymy being the most important one.
- RussNet consists of **4 interrelated files** for basic POS: nouns, verbs, adjectives and adverbs. So far we dealt only with 3 of them, but later we are going to add adverbs as well.
- Each of the 4 files contains a number of hyperonymy trees, with concepts of top levels constituting so called **Top Ontology**.
- Now, we are elaborating mainly internal structure of Russian wordnet and are not dealing with **Inter-Lingual-Index** (ILI).

2. Synset Formation

There are two different ways to define synonymy:

- in terms of substitution
- in terms of semantic similarity.

Although in EWN the weaker notion of synonymy is adopted: «two words are synonyms if there is a statement (class of statements) in which they can be interchanged without affecting truth value», we have to combine substitution method with that of semantic similarity. The reason for such a decision is as follows: in Russian there are many words which are not interchangeable in a context because of the syntactic, stylistic, expressive differences, but they are considered by native speakers as having similar meanings, denoting the same objects, entities, etc., e.g. aspect opposition for verbs.

There are two types of synonymy dictionaries for Russian:

- New Explanatory Dictionary of Russian Synonyms (Apresjan et al.) is following the substitution strategy. The first issue of this dictionary was published in 1999, but so far it includes 132 entries only.
- Dictionary of Russian Synonyms (Evgenjeva,1970) & Explanatory Dictionary of Russian Verbs (Babenko, 1999) are based on semantic similarity.

Unfortunately, conventional Russian lexical resources may be used only partially because they don't cover all the lexis, the words definitions provided are made according to inconsistent patterns, and they may even obscure real semantic relations between words. That's why we can't simply import the data from those resources into RussNet without correcting it by means of our own lexical research procedures.

We begin with the collection of word senses for particular semantic groups of Russian words such as emotional verbs, nouns denoting the social relations and so on. The words realising the hyperlexeme sense were picked out from the sample of fiction or newspaper texts. A mean sample size ranges from 200 to 400 thousand word occurrences, from which about 150 core words and 70 peripheral words with appropriate senses were usually chosen. Having examined the synonymic relation in such groups we saw that words with the most abstract sense were encountered with relatively higher frequency and they would have synonymic equivalents. The hyponyms of the group were rare and may have derivational synonyms, but quite a few synonyms with different roots. So the collected words may be considered to be dominant representatives for respective synsets. Afterwards, extending the sample size or using synonymic information given in a conventional dictionary, we may expand synonymic sets with extra members.

3. Problems and discussion

3.1. Derivation

The Russian vocabulary, in particular verbs and nouns, is characterised by the high degree of derivation motivation. For example, dealing with verbs of thinking in Russian and English, we can see that there is about dozen of verbs with different roots in English (*to think, to contemplate, to consider, to regard, to reflect, to muse, to ponder, to cogitate, to meditate, , to conceive, to imagine, to picture etc*), and only 3 such items in Russian (*думать, мыслить, мозговать*), with a number of affixed derivatives amounting to 30 resultant verbs. Thus the total number of lexemes in Russian may be twice as much as that in English, while the situation with roots may be quite the opposite (Mitrofanova, 1999). From the point of view of frequency this causes specific distribution of lexical items in texts: it is rather flat in English in comparison with Russian sharp peak of frequencies for a hyperonym of this group *думать (think)* see Table 1, Table 2.

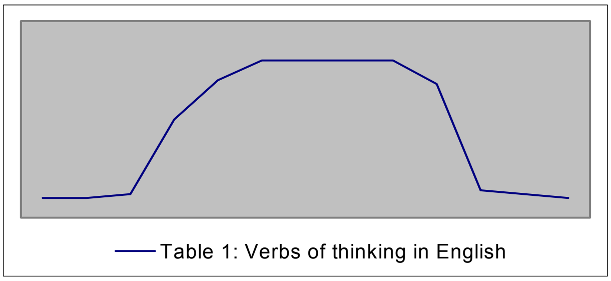


Table 1: Verbs of thinking in English

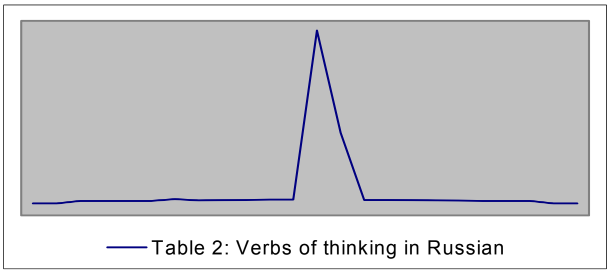


Table 2: Verbs of thinking in Russian

In many cases semantic relations between stem word and its derivatives couldn't be treated in terms of EWN Language Internal Relations (Vossen, 1998). They are more complicated: the main difficulty is that they are relations **between lexical items**, not synsets. Other reasons why we have to introduce new links are as follows:

- There are many almost **unlimited** derivational chains: verb denoting process => noun denoting the process => attribute denoting the relevance to the process => adverb denoting the changing quality and so on, e.g. *удивлять (to astonish, to surprise) - удивление (astonishment) - удивленный (surprised) - удивленно (surprisingly).*
- The important traits of these chains are, that derivatives may be used freely in **paraphrases**: the motivating item may substitute the motivated ones in syntactic transformations. For example, a Russian noun *проверка* (*a check*) is paraphrased as a denotation of the process expressed by Russian verbs *проверить, проверять, провериться, проверяться (to check, to be checked).* These links may be useful for syntactic analysis.
- **Lexical meaning** of derivatives is determined by that of the stem word.
- We would like to stress that verbal nouns inherit also the **syntactical** features of the motivating words. So if we describe the complex system of verb valences, they would be reproduced with little (and well known) changes by nouns denoting the same action or quality, on the one hand, and participants of action, on the other hand.

In those cases when it is possible we regard derivational relations in terms of LIR:

- **SYNONYMY -** relations between words which have the same root and different sets of affixes. They are not expressive and their senses differ so slightly that not every native speaker (researcher) is able to explain the distinction between them. Those words are also rarely interchangeable in the same context. *Семья – семейство (family), зло (malice) – злоба (malice, anger) – злость (malicious anger), бунтарь – бунтовщик (rebel, insurgent, mutineer, rioter), беда (misfortune, calamity) – бедствие (calamity, disaster).*
- **NEAR_SYNONYMY -** relations between
  - ➢ verb and abstract nouns, denoting processes of the same nature, e.g. *двигаться => движение (move => movement)*,
  - ➢ adjectives and abstract nouns, denoting characteristics and qualities, e.g. *красный => краснота (red => redness)*,
  - ➢ adjectives and nouns, e.g. гриб => грибной (fungus => relative to fungi)
  - ➢ verbs and adjectives, e.g. гнить => гнилой (rot => rotten).

In other cases we have to introduce a set of Derivational analogues of LIR, such as:

- **DERIVATIONAL_SYNONYMY** – relation between neutral words and their expressive derivatives. As those words differ from their stem word in style, they are not interchangeable in context, e. g. *старик (old man) => старикан, старикашка* (*impolite appeal to an old man), дом (house) – домик (house to which the speaker has positive emotions)*. Here we follow the idea, offered in Czech WordNet, of special attributes introduction. Thus *домик* will have X_EXPRESSES_ POSITIVE_EMOTION, while *старикашка* – X_EXPRESSES_IMPOLITE .
- **DERIVATIONAL_HYPONYMY** – verb-to-verb, noun-to-noun, adjective-to-adjective relations of following types. For verbs we may use
  - ➢ specific attributes X_HAS_INCHOATIVE or X_HAS_SPECIFIED_DURATION for actions restricted in time duration (inchoatives), e.g. *петь => запеть (to sing => to begin to sing), сидеть => посидеть (to sit => to sit for a while), сидеть => просидеть (to sit => to sit for a long time)*;
  - ➢ an attribute X_HAS_SPECIFIED_RECURRENCY for actions repeated only once or several times, e.g. *кричать => крикнуть, покрикивать (to shout => to shout out once, to shout not aloud many times);*
  - ➢ an attribute X_HAS_SPECIFIED_NUMBER for actions, having many objects involved, e.g. *думать - раздумывать (to think - to ponder about many things for a long time), резать - вырезать (to cut - to cut out some part from many things)*, and so on.

These special verbal derivatives interacting in a complex manner with an aspect category of verbs and having semigrammatical nature. We still don't know in which manner to treat them, on the one hand, aspect pairs look like very close synonyms, though on the other hand, they realise a very important semantic opposition, such as activity ⇔ action. We may introduce specific attributes, as follows: X_HAS_IMPERFECT, X_HAS_ PERFECT.

- ➢ For nouns and adjectives we may add attributes X_IS_SMALL and X_IS_BIG, and possibly several others, when the clear sense component is added by some affixes to the stem word meaning, and the resultant word couldn't be regarded as purely expressive variants; this why we should treat such pairs as *стол => столик (table => small table), дом => домишко (house => small house), пожар => пожарище (fire => big fire), громадный => громаднейший (huge => very huge)* as derivational hyperonym - hyponym.

We should note that the majority of these derivational variants doesn't belong to the core of Russian lexis because of their infrequency in texts. However, the highly inflected nature of Russian may turn any potential derivative into common and frequently used one, that's why all derivational regular models should be taken into account.

Moreover, we may find several cases when an expressive shade may disappear, then a word would change expressive synonym status for a synonym position. Another example of extending the sphere of usage for diminutives may be seen in the Russian spoken language (usually by women), when these words function as oral equivalents for their neutral motivating counterparts, so we may expect that in future they have a chance to become colourless synonyms.

Expressive synonyms and hyponyms may exist beyond the derivational scope, but in these cases they are rather few, irregular, and disputable, that's why it would be adequate to include them into the synset with a proper attribute.

- **DERIVATIONAL_ROLE_RELATIONS** are established to link a verb to its derivatives, designating action participants, such as ROLE_DERIVED_AGENT, ROLE_DERIVED_ OBJECT, ROLE_DERIVED_INSTRUMENT, ROLE_DERIVED_ LOCATION and so on, e.g. *сеять => сеятель, сеянец, сеялка (to sow => sower, seedling, seeding-machine*). The link in the opposite direction is a realisation of the semantic link INVOLVED_IN_ACTION. We are inclined to treat such cases as a specific derivational relation because the semantic link usually has wider scope, e.g. *принимать => приемник (receive => radio set = receiver)*, the object is involved in the first place into the situation *слушать (listen)*. This is usual for complex activity nomination, which as a rule is designated with regard to one action varying from one language to another, e.g. *шить => швея (to sew => seamstress)*. Above we have mentioned the inheritance of syntactic features, moreover, the collocation restrictions of stem verbs may be inherited by their derivatives.

3.2 Adjectives in RussNet

As there is no common solution for treatment of adjectives in EWN, we offer the following one.

We comply with the idea of GermaNet to make use of hyponymy relations wherever it is possible, but our German colleges determine hierarchical structure of adjectives according to semantic fields, while we regard adjectival hyperonymy in terms of their collocations with nouns. We received preliminary results which prove that on the level of adjectives grouping and nouns tree hyperlexeme, it is the **adjective** in Russian that **predicts** certain type of **nouns to collocate with** it, and not vise versa. For example, meaning of *долговязый (lanky)* involves the pointer to a human being, i.e. it can collocate with such nouns as *мальчик (a boy), человек (a man), папа (a father).*

We are prone to the opinion that **adjectival hyponymy trees** can be built according to their collocation with nouns from different levels of hyponymic tree. For example, lets take two adjectives, which express the similar semantic quality – denotation of *height*. In case when one adjective – *высокий (tall)*– may collocate with all nouns denoting "entity": objects, animals, humans and so on, while the other – *рослый (well-grown, srapping)* – collocates only with a certain part of the tree – human beings, the first one may be thought as hyponym for the second one. So checking the co-occurrence of adjectives with nouns, we are to produce hyponymic structure for groups of adjectives denoting the similar quality.

3.3. Verb Valencies

It is generally accepted that syntactic features of words, especially verbs, are determined by their semantic properties, that the meaning of a verb outlines the form and semantic features of words accompanying it.

The semantic and syntactic structure of verb arguments is called the **valencies frame**. Valencies may be thought in terms of morphological noun forms, which are obligatory or optional. This characteristic is vital for Russian syntax, as well as for that of other Slavonic languages (Pala, Sevecek, 1999).

Verbs have different valencies frames associated with dfferent meanings, cf.

- ➢ *Бить (посуду) [to crash]*
- ➢ *Бить (в барабан) [to bit into]*
- ➢ *Бить (врага) [to fight against]*

The minimal form of valency description implies the noun case specification, often it needs the indication of a preposition (or number of prepositions).

We may fix the **semantic** features of nouns as well, which a verb can take as arguments in a sentence. It means we want to use top-level concepts, to deal with **classes** of words, including verb-to-class relations in the synsets. In the example above, the argument of a verb in the first frame is a fragile object, in the second – musical instrument, more precisely – percussion, in the third – human being*s,* military units and so on. This references to the hyponymic tree structure of nouns would be very helpful for syntactic description as well, though sometimes this relation may be very comlicated.

The situation with valencies frames is not clear due to versatility of syntactic preferences of verbs included into a synset, while sometimes they behave uniformly. We'll use **a list of valencies frames** for a synset specifying which one fits the member of a synset. The set of frames is better than separate verb description, because in this case the paradigm influencing the native speaker is presented.

Moreover, it would be very useful to represent the inheritance of syntactic frames of a hyperonym by its hyponyms, e. g. *двигаться (to move) ==> идти (to walk):* hyperonym *двигаться* has valencies frames: (a) "starting point – location", (b) "destination point – location", which are inherited by its hyponym *идти.*

4. Definition Generation

4.1. Subset Sense Definition

We still don't speak about definition generation procedure, but it's vital to have in mind guidelines for definition formulation because dictionary ones for a long time have been a target for an extensive criticism. In this respect we propose several key notes.

4.1.1. Hyponymic Definition

The definition of a synset incorporated into the hyponymic (or troponymic) tree should be constructed on the following pattern "the dominant **lexeme** of the **hyper** level **plus** a **distinguishing part** showing difference between co-hyponyms", e.g. *плыть (to swim)* has hyperlexeme: «to move in certain direction» + differentiation: «on the surface or in depth of water using special organs», *лететь (to fly)* has hyperlexeme: «to move in certain direction» + differentiation: «in the air using wings». In this case there is no Russian hyperlexeme denoting *moving in some direction*, though it's important to oppose this way of moving to the other one in various direction, with repetitions, to and fro.

It's clear that in case of a large number of co-hyponyms the problem may become practically insolvable because of a great number of necessary differential features, then it would be better to use other types of defining or artificial names (used in GermaNet) uniting several lexemes into a cluster.

4.1.2. Meronymic Definition

The definition of a synset incorporated into the meronymic relations may be based on either holonym, or meronym.

In the first case, a holonym is the referential part of the definition (similar to hyperlexeme), but a simple indication that something is a part of the holonym is not sufficient, so it is usually supplied with a special function (for artefacts) or construction peculiarties. For example, structure «part + construction characteristic + holonym + function» may be used: *крыша (roof)* = «the upper part of the building, covering it from precipitation».

In the second case, a limited number of meronyms may be used for generation of list-type definition, e.g. *фигура (chessman)*: «king, queen, castle, knight, bishop in chess opposed to pawns».

4.1.3. Derivational Definition

In those cases when a synset is associated with a purely derivational link we use a definition describing the additional sense of the derivational affixes, e.g. *столик* «a small table», *генеральша* «general's wife».

4.1.4. Semantic Pointer Definition

The simplest way of defining the quality is to show the synonyms expressing it, which are united in the synset, so in this case we have a rudimentary definition equal to an ordinary synset. This type of definition is frequent for adjectives and adverbs.

Antonymic definition is adequate in those cases when one member of the antonymic pair is marked showing the positive content while the other shows its absence, e.g. *глупый (foolish)* «not clever» <=> *умный (clever)* «having the intellect».

Causative definition is alike the derivational one so as it makes implicit the causative copula and the final state of transition, in Russian there is a specific affix with anticausative meaning, e.g. *поднять (raise)*: *каузировать подняться (cause to rise)*. Usually in such a definition the artificial causative is used, which is the transliteration of English *cause*, because a Russian equivalent *заставить* means 'to enforce', that is not neutral at all.

Moreover, using semantic attributes, such as X_HAS_IMPERFECT, X_HAS_PERFECT, X_IS_SMALL, X_IS_BIG etc., incorporated into the WordNet structure, we may later elaborate a procedure for automatic definition generation.

4. Conclusions

To sum up we may say that RussNet presently covers the core of the Russian lexis (the resulting number of synsets is more than 2500). So it can be regarded as a reliable starting point for further extending and elaboration of the system, which will be carried out by addition of peripheral groups of words, emotionally coloured lexis and derivatives, in particular. This should enrich the content of the database. The introduction of new relations allows us to perform more adequate semantic analysis of the Russian language.

5. References

Apresjan, U. (ed.) (1997). Новый объяснительный словарь синонимов русского языка (=New Explanatory Dictionary of Russian Synonyms). Moscow.

Babenko, L. (ed.) (1999).Обяснительный словарь русских глаголов (=Explanatory Dictionary of Russian Verbs). Moscow.

Evgenjeva, A. (ed.) (1970). Словарь синонимов русского языка. (=Dictionary of Russian Synonyms) (vol 1-2). Leningrad.

Miller, G. et al (1993). Five Papers on WordNet. Technical Report, Cognitive Science Laboratory, Princeton University. ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps

Mitrofanova, O. (1999). Структурный анализ сигнификативного значения: на материале глаголов процесса мышления английского и русского языков (Structural Analysis of Sense: Verbs of Knowing in English and Russian). PhD thesis. St-Petersburg State University, Philological Faculty, Department of Applied, Structural and Mathematical Linguistics.

Naumann, K. (2000). Adjectives in GermaNet. http://www.sfs.nphil.uni-tuebingen.de/Adj.html

Ozhegov, S. (1984). Словарь русского языка (=Dictionary of Russian). Moscow.

Ozhegov, S., Shvedova, N.(1992). Толковый словарь русского языка (=Explanatory Dictionary of Russian). Moscow.

Pala, K., Sevecek, P. (1999). The Czech WordNet, EuroWordNet (LE-8928). Deliverable 2D014. http://www.hum.uva.nl./~ewn/docs.html

Vossen, P. (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht: Kluwer.

Словарь современного литературного русского языка.(1991). (=Dictionary of Modern Literary Russian) (vol. 1-17). Moscow-Leningrad.

Development and Use of Thesaurus of Russian Language RuThes

Natalia V. Loukachevitch* and Boris V. Dobrov*

Research Computing Center of Moscow State University
339, Research Computing Center of Moscow State University,
Vorobyevy Gory, Moscow, 119899, Russia
{louk, dobroff}@mail.cir.ru

Abstract

In the paper we describe the main principles of developing Thesaurus of Russian Language RuThes, which is constructed specially as a tool for automatic text processing. The thesaurus contains more than 95 thousands words and multiword expressions. It has a specific system of conceptual relations, describing existential properties of concepts. Means of description and disambiguation of lexical ambiguity are discussed. The technology of development the bilingual resource based on RuThes is described. We also consider current stage of the thesaurus and describe the use of the Thesaurus in various applications of automatic text processing.

1. Introduction

Large volumes of electronic text collections require mighty tools for their processing. Texts in these collections include thousands of various words and syntactic constructions, they can have various sizes and styles. All these factors pose an important question what linguistic resources facilitating processing large collections of electronic documents could be.

The paper is devoted to description of main principles of development of the Thesaurus of Russian Language RuThes, which belongs to the same type of such linguistic resources as WordNet (Miller et al., 1990) and EuroWordNet (Climent et al., 1996).

This work arises from our experience in creation of domain-specific Thesaurus on Sociopolitical Life, which was constructed as a tool for automatic conceptual indexing in the large domain of social life (Loukachevitch et al., 1999). Development of Sociopolitical Thesaurus for automatic text processing of large text collections required use of two different traditions: the tradition of development of information-retrieval thesauri for manual indexing, which pay specific attention to terminology and representation of domain-specific relations (LIV, 1994; UNBIS, 1976; EUROVOC, 1995), and the tradition of development of linguistic resources with their attention to description of single words, lexical ambiguity, semantic relations.

The Sociopolitical thesaurus was used in such applications of automatic text processing as term disambiguation, automatic conceptual indexing, knowledge-based text categorization, automatic text summarization (Loukachevitch et al., 1999). The Sociopolitical thesaurus is an information retrieval tool in University Information System RUSSIA (Russian inter-University Social Sciences Information and Analytical Consortium; www.cir.ru/eng/).

The technique of text processing using Sociopolitical thesaurus is based on lexical cohesion property of coherent texts, that is, the thesaurus relations were used to find semantically related sets of terms in texts (Loukachevitch & Dobrov, 2000). For several years the results of the text processing were tested through manual analysis. We tried to understand how thesaurus relations work in the thematic structure of coherent texts. This activity led us to development of RuThes, a linguistic resource for automatic text processing of large Russian text collections.

Now thesaurus RuThes includes 95 thousand Russian words (nouns, verbs, adjectives), expressions and terms, 105 thousand senses, 42 thousand concepts (synsets). In contrast to European wordnets we began to describe Russian-English relations of RuThes after considerable part of RuThes had been already created.

2. General Structure of RuThes

RuThes is a hierarchical net of concepts. Every concept has a set of its textual expressions, a synonymic row (a synset in terminology of WordNet) and a set of relations with other concepts of the thesaurus. So its general structure is the same as the structure of WordNet and EuroWordNet.

RuThes consists of two main parts: general lexicon and Thesaurus on Sociopolitical life (Figure 1).

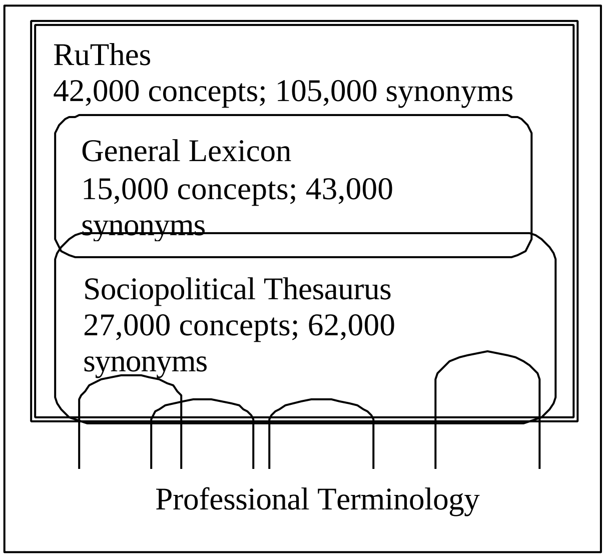

RuThes
42,000 concepts; 105,000 synonyms
General Lexicon
15,000 concepts; 43,000 synonyms
Sociopolitical Thesaurus
27,000 concepts; 62,000 synonyms
Professional Terminology

Figure 1. General Structure of RuThes

Thesaurus on sociopolitical life includes concepts which correspond to a domain of social life, these words and terms are usually thematically significant: *car, river, town, economy, computer, sports, serviceman* and many others. The domain of Sociopolitical Thesaurus is not domain of social research, but comprises situations and problems of social life, which are discussed in official documents and newspapers. Sociopolitical thesaurus encompasses as words and expressions usually included in general explanation dictionaries and terminology of such domains as economy, law, defense and others. Besides Sociopolitical thesaurus includes the geographical subdomain describing 7000 geographical names.

General lexicon contains concepts, which can be met in texts of any domains, for example, *part, create, new*. In texts these words and expressions are less significant, they usually express relations and features of main entities discussed in texts. Besides general lexical contains concepts expressing human emotions, feelings, personal human relationships. General lexicon contains 15 thousand concepts (from 42 thousand in RuThes), 33 thousand words and language expressions (from 95 in RuThes).

The main goal of the division was as follows: this borderline separates very lexically ambiguous area from much less ambiguous, very relational area from much more thematically significant. The result of this division is that Sociopolitical thesaurus is used in various applications of automatic text processing for several years. General lexicon is now under development. Its numerous multiple senses are added and corrected. But Sociopolitical thesaurus and General lexicon are parts of the same system. Therefore if necessary, knowledge from General lexicon is used in computer applications together with Sociopolitical thesaurus.

3. Synsets in RuThes

3.1. Description of different parts of speech in RuThes

Elements of synsets in RuThes (below Thesaurus) are: single words (nouns, verbs, adjectives), noun groups, verb groups, adjective groups. We describe semantically equivalent words belonging to different parts of speech as elements of the same synset - our tradition from 1994.

Here and below we present examples in English to show how our decisions would look in English. So a synset looks like: *partake, participate, participation, participative, participatory, take part.*

Every synonym has its part of speech tag, and this information can be used in automatic text processing.

Incorporation of parts of speech makes description of relations more consistent. If we create three different concepts for different parts of speech and know about their semantic equivalence, we have to repeat similar relations for them. This leads to inconsistency in description of relations. For example, in WordNet 1.6. we can see that word *engagemen*t is in the same synset as *participation*, but there is no relation between *engage* and *participate.* However in Webster (1999) we can find the following example: *to engage in business or politics*, which means that the relation between verbs has to be described.

This incorporation also means that the hierarchy, the set of relations is the same for all parts of speech.

3.2. Multiword expressions in RuThes

We pay special attention to description of multiword expressions and terms as sources for lexical disambiguation and representation of situational and encyclopedic knowledge. A number of multiword expressions in the Thesaurus is 42 thousand (of 95 thousand).

We began development of Sociopolitical thesaurus using semi-automatic methods to find multiword terms in text collections of official documents and newspaper articles. Our procedure of terms acquisition consisted of two stages. At the first stage term-like expressions were automatically identified in the texts of the corpus. Rules defining term-like expressions included syntactical and lexical conditions. At the second stage our specialists had to look through the revealed expressions, choose terms from them and add new terms to the Thesaurus (Lukashevich, 1995). The procedure was working during four years, processed more than 200 Mb of texts and collected more than 200 thousand term-like expressions. It was stopped because it became difficult to find new useful terms, terminology coverage became very high.

From that experience we understood how important to add unambiguous multiword expressions containing ambiguous words to conceptual synonymic rows. They diminish percentage of ambiguous words in a text and help disambiguate neighbor expressions. Since we specially seek unambiguous multiword expressions in any sources we have: in glosses and examples of dictionaries, in text collections. For example, the following expressions could be added to synonymic rows of WordNet and are very useful in automatic text processing:

> *Petition to god (*sense2 of noun *petition);*
> *transfer to private ownership* (verb *privatize),*
> *conductor of an orchestra (*sense1 of noun *conductor)*

As an example how a full synonymic row of multiword expressions could looks, let us see the synonymic row of Russian concept *ZDRAVOOHRANENIE* (*PUBLIC HEALTH*), which is similar to the following English list:

> p*ublic health, community health, health care, health care sector, health care system, health field, health of population, health promotion, provision of health, public health field.*

So one can see how this list diminishes necessity to disambiguate such “difficult” words as *care, sector, field, system, public.*

A multiword expression can also initiate a new concept. There are several factors that can make possible creation of a new concept based on a multiword expression:

- A multiword expression presents an important and frequent enough subtype of a concept already described in the Thesaurus;
- A multiword expression is unambiguous and contains very ambiguous words;
- A multiword expression has conceptual relations that do not follow from its constituent parts;
- A multiword expression has relations with concepts of lower levels, based on single words, so a new concept additionally structures the thesaurus knowledge, can join separate conceptual substructures of the thesaurus net.

3.3. Name of concept

Every concept of the thesaurus has a unique name, which has to be clear and unambiguous for native speakers. Name of a concept can be

- one of unambiguous synonyms;
- a multiword term which is unambiguous and possible as one of textual expression corresponding to a concept;
- a pair of synonyms;
- a synonym with a fragment of the definition of a concept.

This name presents the whole synonymic row in different representations of results of text processing, for example, in structural summary of a text which is very convenient in cross-language information retrieval (Loukachevitch & Dobrov, 2000) or as explanation means for knowledge-based text categorization systems.

A concept usually does not have a full gloss but formulation of its name has to be enough to find a corresponding sense in explanation dictionaries if necessary.

4. Description of lexical ambiguity

In linguistic resources intended for automatic text processing there is a serious problem how detailed division of senses must be. The sources of the problem are as follows:

- it is difficult to disambiguate close meanings during automatic text processing in large domains;
- it is impossible to refine query with help of a user because a user must not understand and distinguish subtle linguistic distinctions;
- at last close meanings (even if we have divided and can disambiguate them) are often both relevant or not relevant to a query.

Therefore we have to understand, what types of ambiguous terms it is necessary to distinguish and represent as different concepts of the Thesaurus.

In a linguistic resource represented as a conceptual net the desire to reduce number of senses is in contradiction with other problem: if two senses have different sets of conceptual relations (especially different sets of links to lower levels of the conceptual net), then their clustering can lead to loss of descriptive clarity and new problems in efficiency of automatic text processing.

Therefore in RuThes we do not cluster senses that have different hyponyms and/or parts. If the difference between sets of conceptual relations consists only of hypernyms, sense clustering is possible.

For example, it is impossible to cluster concepts corresponding to the senses of word *building* as process and result as proposed in (Pustejovsky, 1995), because in the Thesaurus difference in conceptual relations between the concepts is significant. Compare fragments of lower levels corresponding to these concepts:

CONSTRUCTION OF BUILDINGS (build, building, building construction, construct, construction….)
    RESIDENTIAL CONSTRUCTION (home construction, homebuilding, home building…)
    ...
    COSTRUCTION EQUIPMENT(building equipment)
        TOWER CRANE
        BULLDOZER
        EXCAVATION EQUIPMENT
    ...

BUILDING…
    PUBLIC BUILDING
        ADMINISTATIVE BUILDING
        MUSEUM…
        SCHOOL BUILDING…
    ...
    RESIDENTIAL BUILDING
        APARTMENT HOUSE
        VILLA
    ...

The problem of description of close senses became less serious if it is possible to describe relations between corresponding concepts. The relations of RuThes allow us to connect various types of polysemic senses, and in automatic processing if it was not possible to distinguish a correct meaning, the most broad concept among all related senses is chosen in default way. In general, it is possible to have a special indicator, showing which concept can be chosen in default.

For example, we can introduce two concepts *SCHOOL (EDUCATIONAL ORGANIZATION)* and *SCHOOL BUILDING*, connect them with relation WHOLE-PART and include word *school* in synsets of both concepts. In automatic text processing if it is not proven that a school building is discussed, concept *SCHOOL (EDUCATIONAL ORGANIZATION)* is chosen. It means that there is no real difference between description of these polysemic senses as two concepts or a single concept. In RuThes choice of separated or clustered description of close polysemic senses depends on if sense concepts are central in the thesaurus net and require their own sets of lower relations or they are peripheral.

5. Relations in RuThes

Linguistic resources intended for automatic text processing usually include descriptions of semantic relations between their entries such as ‘part’, ‘agent’, ‘material’, ‘time’, ‘cause’, ‘result’, and others. At the same time when huge conceptual-based resources are developed, it is supposed that these resources have to be used in automatic text processing of large and heterogeneous text collections. However, at present text processing systems can not provide deep linguistic analysis of such texts. It means that a computer system can not check if described relations are valid in a current text. Therefore other, not semantic, characteristics of any relation become especially important, if a relation can change or disappear in a specific situation described in a text. These characteristics can be considered as existential characteristic of a relation.

Therefore if we describe that a tree is a part of a forest, but in fact a tree can grow in a lot of other places, the system can not rely on this description because in a specific text the relation can be not valid. It can lead to problems in efficiency of automatic text processing.

To test changeability of a relation between concepts *C1* and *C2* it is necessary to answer the following questions:

1) if every example of a concept *C1* has the relation with an example of a concept *C2* (and vice versa);
2) if an example of concept *C1* has the relation with *C2* (or its example) during all time of its existence, for example, concept *GARMENT* can be considered as *CONSUMER GOODS* (as described in WordNet 1.6), but when a specific person wears garment, it ceases to be goods;
3) if all properties of a concept *C1* are properties of concept *C2*, for example, concept *SHIPWRECK* loses very important properties of concept *SHIP*.
4) if existence of a concept *C1* is impossible without existence of concept *C2* or existence of an example of a concept *C1* is impossible without an example of concept *C2* (dependency relations (Guarino, 1998)), for example, existence of concept *BOILING* is impossible without existence of concept *LIQUID*.

At present description of relations in RuThes do not present semantic nature of relations distinct from hyponymy-hyperonymy relations and part-whole relations, but its existential properties. At the same time it gives additional very powerful possibility not to decide what a semantic name of a relation can be. It is very important for complex relations such as *CREDITOR – BANCRUPCY*: if the name of the relation is ‘agent’ or ‘source’ or both.

Current names of conceptual relations in RuThes were introduced in earlier version of Sociopolitical thesaurus and arise from names of relations in conventional information retrieval thesauri. There are three basic relations:

1) BT-NT relations (broader-narrower terms) is now used as equivalent to hyponym-hypernym relations;
2) WHOLE-PART relations for descriptions of conventional parts, properties and participants of situations;
3) RT (related term) relations for description of all other relations, which can be symmetrical and nonsymmetrical. Nonsymmetrical RT relation is denoted as RT1 – RT2 and serves for description of dependency relations.

Let us see fragments of description of concept *RIVER* to see usage of PART and RT relations

RIVER

PART	*RAPIDS OF A RIVER* *(Russian ‘bistrina’)*
PART	*WATERFALL* *(Russian ‘vodopad’)*
PART	*MOUTH OF A RIVER* *(Russian ‘ust’e’)*
RT1	*FRESHWATER* *(‘presnaya voda’)* /* concept *RIVER* does not exist without existence of concept *FRESHWATER* therefore there is a dependency relation denoted as RT1. At the same time a lot of concepts depend on existence of concept RIVER. So below reverse relation RT2 is used */
RT2	*CATCHMENT BASIN* (Russian ‘bassein reki’)
RT2	*HYDROELECTRIC PLANT* (‘gidroelectrostancia’)
RT2	*EMBANKMENT* (‘nabereznaya’)
RT2	*BOTTOMLAND* (‘poima’)
RT2	*RIVER TRANSPORT* (‘rechnoi transport’)
RT2	*SLUICE GATE* (‘shljuz’)

If for a BT or WHOLE relation there is an answer ‘OFTEN’ to one of questions 1-3, then a relation is marked with special modifiers.

If a relation can be considered as a default relation or there are only two main alternatives, we mark the relation with modifier V (variability)

If a relation exists during most time of existence of an example of a concept, we mark a relation with modifier A (aspect, point of view). The same modifier is used if a relation does not preserve all properties of an upper concept. For example:

PENSIONER

BT $_{\text{V}}$	*OLDER PERSON*
BT $_{\text{V}}$	*DISADVANTAGED PERSON*
WHOLE $_{\text{A}}$	*PENSION SYSTEM*

So we described that a pensioner is often an older person and a disadvantaged person. A pensioner is a role in pension system, which does not characterize it fully because of two first relations. In fact, a pensioner is also a social status. Therefore if a text mentions pensioners, it does not always mean that the text discusses some problems of pension system.

Every type of conceptual relations has its own set of properties such as transitivity and inheritance. Modifiers restrict transitivity of relations (Loukachevitch & Dobrov, 2002).

6. Lexical coverage of RuThes

Now thesaurus RuThes includes 95 thousand Russian words (nouns, verbs, adjectives), expressions and terms, 105 thousand senses, 42 thousand concepts (synsets). At present we have finished comparison of lexical units in RuThes and in a text collection of more than 600 thousand documents (Russian official documents and newspaper articles). Analysis of 100,000 most frequent lemmas of the collection (frequency > 25) showed that about 7 thousand lemmas are necessary to describe in RuThes. We plan to continue study of the text collection and to add new lexical units in RuThes for next 100 thousand lemmas (frequency > 10). We suppose that this stage will give us other 5-7 thousand words to include in RuThes.

The lexical analysis of the collection allows us to describe new words, not included in contemporary Russian dictionaries, and see new usage of words that are considered in the dictionaries as obsolete.

Other important stage of our current work is verification of sense representation for polysemic and homonymic words in RuThes. Beginning from very frequent words we analyze senses of every lexeme described in various dictionaries of Russian language (Ozhegov & Shvedova, 1999; BTS, 1998) and decide if

a) all senses of a lexeme have to be represented;
b) there are obsolete senses;
c) different senses can be represented as a single concept;
d) a sense is only used within multiword expression.

So current stage of development of RuThes can be characterized as verification and correction.

7. RuThes and English linguistic resources

Development of cross lingual linguistic resources is a very important task. For Russians bilingual text processing of Russian to English and English to Russian is especially significant. We began development of RuThes from Sociopolitical thesaurus, which is an important searching tool in our information system. To provide bilingual retrieval in our information system we began to develop Russian-English Sociopolitical Thesaurus. It means that we could not connect RuThes and WordNet because of absence of significant in our technology concepts of the sociopolitical domain in WordNet. Besides we considered collection of multiword terms as very important for any language. The following list presents English terms included to English part of Sociopolitical Thesaurus recently and not included to WordNet: *wheelchair user, construction area, airline ticket, travel field, home building, civil rights activist, top manager, produce market, cargo shipper, stress disorder* and others (terms are extracted from newspapers).

Development of bilingual Sociopolitical thesaurus has the following main stages.

At first Russian terms were translated into English using traditional bilingual dictionaries (Apresyan & Mednikova, 2000; Multilex, 1996). We received 30 thousand terms in the English part of our Thesaurus. However these translation could not provide rich synonymic rows we needed and could not provide terms describing phenomena that are absent in Russia but are significant for other countries.

Therefore at the second stage we took well-known American and British dictionaries and thesauri: Webster dictionary (1999), Longman dictionary (1995), Collins (1990), WordNet (Miller et al., 1990), Thesaurus Roget's (1991), information retrieval thesauri Legislative Indexing Vocabulary (LIV, 1994), EUROVOC (1995), UNBIS (1976)). Our specialists analyzed these resources and manually extracted terms contained in these resources as vocabulary entries, parts of explanations, examples.

Therefore an English expression can have a mark, indicating its origin. For example, a concept *EQUALITY BETWEEN MEN AND WOMEN* has the following synonymic expressions:

equal rights for women (WordNet's gloss)
equal rights of men and women (EUROVOC)
*equality between sexes (*Multilex*)*
*equality between women and men (*texts - documents of Council of Europe)
gender equality (texts)
sex equality (texts*)*.

Text variants of related concept *SEX DESCRIMINATION* are as follows*:*

*Discriminations on the ground of sex (*texts)
*Gender descrimination (*LIV*)*
Sex discrimination (LIV*)*
*Sexism (*Webster, WordNet*)*

This stage is planned to take two years and be finished before 2003. Now the English part of Sociopolitical thesaurus comprises 48 thousand English terms.

Now we began the third stage of the development – revision and correction of collected material.

And the fourth stage is use of the bilingual resource in various applications of automatic text processing, which will lead to further improvement and enrichment of our linguistic resource.

It is important to stress that during analysis of dictionaries our specialists were approved to make Russian-English connections for any Russian words in RuThes (not only from Sociopolitical Thesaurus). Full volume of included English words and expressions is more 62 thousand entries, 67 thousand senses. So this work can be considered as a significant basis for connection to other English structural resources.

8. Use of RuThes in text processing applications

8.1. Use of Sociopolitical thesaurus

Thesaurus on sociopolitical life is used in automatic processing applications since 1996. The Thesaurus is a searching tool in University Information System RUSSIA (UIS RUSSIA, www.cir.ru/eng/), containing more than 600 thousand documents. The text collection of this information system includes such various types of documents as official documents of Russian Federation, legislative acts, international treaties, newspaper articles and statistical reports.

The Sociopolitical thesaurus is used as a linguistic resource in such information retrieval applications as automatic conceptual indexing, knowledge-based text categorization, automatic text summarization (Loukachevitch et. al., 1999). In these applications a thesaurus-based technique of construction of thematic representation of texts is used (Loukachevitch & Dobrov, 2000).

In (Loukachevitch & Dobrov, 2002) we describe an experiment which showed that use of this part of RuThes in information retrieval was much more efficient than retrieval based on vector model. Average precision of document retrieval with the Sociopolitical thesaurus (using its synonyms and hierarchy) was 1.4 times more than average precision of vector retrieval.

8.2. Use of RuThes in text categorization systems

RuThes is currently used as a linguistic resource for knowledge-based text categorization systems.

There are a lot of applications where machine-learning approaches (Joachims, 1998) to text categorization are impossible to use. There can be no sufficient training collection, or a system of categories can include hundreds of hierarchical categories. In these cases a knowledge-based technique using RuThes can be appropriate (Loukachevitch, 1997). Knowledge described in RuThes substitutes information received from training examples in machine learning approaches.

In our text categorization technique the categories are manually described using Boolean expressions of a relatively small number of 'supporting' concepts. Boolean expressions including all necessary concepts of RuThes are generated on the basis of properties of the Thesaurus relations. The resulted Boolean expressions usually include much more disjunctive and conjunctive components, sometimes in hundreds times more. It became possible owing to detailed presentation of various aspects of described concepts and careful testing of the Thesaurus relations.

One of our last text categorization systems categorizes Russian legislative documents using the system of 1168 categories (3-4 levels of hierarchy), other text categorization system categorizes public opinion polls (almost 400 categories).

Description of categories in large hierarchical systems of categories usually requires large range of lexical knowledge from very specific terminology to very general words. For example, one of categories for categorization of public opinion polls was “Image of woman” and required detailed descriptions of human traits, the list of which was stored in RuThes.

9. Conclusion

In the paper we described main principles of developing Thesaurus of Russian Language RuThes, which is constructed specially as a tool for automatic text processing. The thesaurus contains a lot of multiword expressions, has a specific system of conceptual relations, describing existential properties of concepts, has specific means for lexical disambiguation. We describe current stage of the Thesaurus developing in comparison to 100,000 the most frequent lemmas of the text collection of University Information System RUSSIA, including more than 600 thousand documents. Now thesaurus RuThes is a basis for development the bilingual Russian-English resource for cross lingual text processing. Also we consider the use of the Thesaurus in various applications of automatic text processing.

10. Acknowledgements

Partial support for this research is provided by the Russian Foundation for Humanities through grant # 00-04-00272.

11. References

Apresyan, Yu.D. and Mednikova E.M., 2000. *Noviy Bolshoi anglo-russkiy slovar.* Yu.D.Apresyan and E.M.Mednikova (eds.), Moscow: Russkiy Yazyk. 5th edition. (*in Russian*).

BTS, 1998. *Bolshoi Tolkoviy Slovar Russkogo Yazyka*. S.A.Kuznetsov (ed.), Sankt Peterburg: Norint (*in Russian*).

Climent, S., Rodriguez, H. and Gonzalo, J., 1996. Definitions of the links and subsets for nouns of the EuroWordNet project. - Deliverable D005, WP3.1, EuroWordNet, LE2-4003.

Collins, 1991. *Collins English Dictionary*. HarperCollins. 3rd edition.

EUROVOC, 1995. *Thesaurus EUROVOC*. Vol.1-3, European Communities: Luxemburg: Office for Official Publications of the European Communuties. 3rd edition. English version.

Guarino, N., 1998. Some Ontological Principles for Designing Upper Level Lexical Resources. In *Proceedings of First International Conference on Language Resources and Evaluation*.

Joachims, T., 1998. Text categorization with support vector machine: Learning with many relevant features. In *European Conference on Machine Learning (ECML-98)*, Springer Verlag, 137-142.

LIV, 1994. *Legislative Indexing Vocabulary*. Congressional Research Service. The Library of Congress. Twenty-first Edition.

Longman, 1995. *Longman dictionary of contemporary English*. Harlow (Essex): Longman.

Loukachevitch, N., 1997. Knowledge Representation for Multilingual Text Categorization. In. *AAAI Symposium on Cross-Language Text and Speech Retrieval*, AAAI Technical Report, 133-142.

Loukachevitch, N., and Dobrov, B., 2000. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems. Machine Translation Review 11: 10-20.

Loukachevitch, N. and Dobrov, B., 2002. Evaluation of Thesaurus on Sociopolitical Life as Information-Retrieval Tool. In *LREC2002 Proceedings*. Las Palmas.

Loukachevitch, N.V., Salii, A.D. and Dobrov, B.V., 1999. Thesaurus for Automatic Indexing: Structure, Developement, Use. In P. Sandrini (ed.), *Proceedings Fifth International Congress on Terminology and Knowledge Engineering*. Vienna: TermNet. 343-355.

Lukashevich, N., 1995. Automated Formation of an Information-Retrieval Thesaurus on the Contemporary Sociopolitical Life of Russia. *Automatic documentation and mathematical linguistics*. 29(2): 29-35.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., 1990. Five papers on WordNet, *CSL Report, 43*, Cognitive Science Laboratory, Princeton University.

Multilex, 1996. *Multilex 1.0a. Anglo-russkiy elektronniy slovar.* Medialingua Ltd.

Ozhegov S.I. and Shvedova N.Yu., (1999). *Tolkoviy Slovar Russkogo Yazyka*. Russian Academy of Sciences. Institute of Russian Language. Ì oscow: Azbukovnik. 4th edition. (*in Russian*).

Oxford, 2000. *The New Oxford thesaurus of English*. P. Hanks, (ed.), Oxford. Oxford Univ. Press.

Pustejovsky, J., 1995. *The Generative Lexicon*. Cambridge, Massachusetts, London, England: The MIT Press.

Roget, 1987. *Roget's thesaurus of English words and phrases*. B. Kirkpatrick (ed.), Harlow (Essex): Longman.

UNBIS, 1976. *UNBIS Thesaurus.* English Edition, Dag Hammarskjold Library of United Nations, New York.

Webster, 1999. *Random House Webster's Unabridged Dictionary*. Version 3.0. Random House, Inc.