Symposium on Lexicography XI

at the University of Copenhagen

Edited by Henrik Gottlieb,
Jens Erik Mogensen and Arne Zettersten

Sonderdruck aus:
LEXICOGRAPHICA
Series Maior 115
ISBN 3-484-39115-4

Max Niemeyer Verlag
Tübingen 2005
Table of Contents

Introduction ................................................................................................................ IX
Acknowledgements .................................................................................................. XIV
Vorwort ....................................................................................................................... XV
Danksagungen ........................................................................................................... XXI

Jean Aitchison
Absolute Disasters: The Problems of Layering ............................................... 1

Herbert Ernst Wiegand
Äquivalenz, Äquivalentdifferenzierung und Äquivalentpräsentation in zweisprachigen Wörterbüchern. Eine neue einheitliche Konzeption .................. 17

Gertrud Greciano
Phraseographische Prioritäten, erfüllt und unerfüllt .......................... 59

Andrea Abel & Vanessa Weber
ELDIT – Electronic Learner’s Dictionary of German and Italian: Semi-Bilingual, Bilingualised or a Totally New Type? ................. 73

Arleta Adamska-Salaciak
A Lexicographical Remake: The Story of the Kościuszko Foundation (English-Polish) Dictionary ................................................. 85

Richard Almind
Dictionaries through Databases ........................................................................ 101

Laurie Bauer
The Illusory Distinction between Lexical and Encyclopedic Information .......... 111

Steven Berbeco
Alphabetizing the Unalphabetical: The Hungarian Sign Language Dictionary Project .............................................................. 117

Henning Bergen Holtz
Falsche und richtige lexikographische Definitionen .................................. 125

Tove Bjerneset
The NORDLEXIN-N Dictionary Project: Part II ........................................... 133

Evelyn Breiteneder & Dmitrij Dobrovolskij
Specific Problems of Text Lexicography (Fedor Dostoevskij and Karl Kraus) .......................... 147

Ulrich Busse
Beobachtungen und Überlegungen zur Aussprache englischer Wörter im Deutschen und ihrer Repräsentation in ausgewählten Wörterbüchern .......... 159
<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timothy Colleman</td>
<td>On Representing Verb Complementation Patterns in Dictionaries: The Approach of the CVVD</td>
</tr>
<tr>
<td>John Considine</td>
<td>“Our Dictionaries Err in Redundancy”: The Problem of Encyclopedism Past, Present, and Future</td>
</tr>
<tr>
<td>T.P. Dolan</td>
<td>The Compilation of a Dictionary of Hiberno-English</td>
</tr>
<tr>
<td>Ken Farø</td>
<td>Korpuslexikographie und Flexionsmorphologie</td>
</tr>
<tr>
<td>Franziskus Geeb</td>
<td>Perspektiven der Rechneranwendung in der Lexikographie: Lexikographische Datenmodellierung und Datenstrukturierung in relationalen Datenbanken</td>
</tr>
<tr>
<td>Peter Gilliver</td>
<td>Materials and Methodologies for the New Oxford English Dictionary</td>
</tr>
<tr>
<td>Maurizio Gotti</td>
<td>The Evolution of English Canting Lexicography in the 16th and 17th Centuries</td>
</tr>
<tr>
<td>Rufus H. Gouws</td>
<td>Multiple Niching</td>
</tr>
<tr>
<td>Gisela Harras &amp; Kristel Proost</td>
<td>The Lemmatisation of Idioms</td>
</tr>
<tr>
<td>Dieter Herberg</td>
<td>Wissen über (neue) Wörter: Ein Internetwörterbuch entsteht</td>
</tr>
<tr>
<td>Lars Holm</td>
<td>Ein schwedischer Calepinus aus dem 17. Jahrhundert</td>
</tr>
<tr>
<td>Anne Lise Laursen &amp; Grete Duvå</td>
<td>Cyberlexicography in LSP: New Aspects of Components and Structures in the Dictionary</td>
</tr>
<tr>
<td>Anatoly Liberman</td>
<td>Scandinavian Etymological Lexicography</td>
</tr>
<tr>
<td>Klaus-Dieter Ludwig</td>
<td>Wortschatzentwicklung und Wörterbuch</td>
</tr>
<tr>
<td>Geart van der Meer</td>
<td>On Taboo and Slang in English Pedagogical Lexicography</td>
</tr>
<tr>
<td>Rosamund Moon</td>
<td>Dictionaries and Metaphor, Metaphor and Dictionaries</td>
</tr>
</tbody>
</table>
Judith Muráth  Wörterbuchbenutzung von Fachübersetzerstudenten. Ihre Erwartungen an ein Fachwörterbuch ....................... 401

Maria José Pereira de Oliveira  Genre Analysis and the Elaboration of Glossaries ............. 417

Gunnar Persson  Dictionaries as Mirrors of Social and Cultural Change ...... 427

Heribert Picht  Die Fachwendung in fachlexikographischen Produkten ....... 435

Tadeusz Piotrowski  An English-Polish Dictionary for Polish Teenagers: A Dictionary by a Non-Native for Non-Natives .................. 443

Santiago Posteguillo Goméz & Jordi Piqué-Angordans  Computer Terminology: Developing an active Bilingual English-Spanish Dictionary ........................................ 453

Stefan J. Schierholz  Valenzwörterbücher für Substantive ................................ 475

Henrik Køhler Simonsen  User Involvement in Corporate LSP Intranet Lexicography ................................................................. 489

Włodzimierz Sobkowiak  Lexicographic Phonetics or Phonetic Lexicography? .......... 511

Laimutė Stankevičienė  English-Lithuanian Lexical Pseudo-Equivalents. A Lexicographical Aspect ......................................................... 521


Andrejs Veisbergs  Ideology in Dictionaries – Definitions of Political Terms ... 537

Kadri Vider & Heili Orav  Estonian Wordnet and Lexicography .................................... 549
Kadri Vider & Heili Orav

Estonian Wordnet and Lexicography

1. Introduction

There has long been a need for computer thesauri in Estonian lexicography. By 1997 it was clear that, in addition to morphological and syntactical analysis, a lexical database based on word semantics was needed. The WordNet (WN) — created by G. A. Miller and others at Princeton University in the 1980s — already existed, and we followed suit. Work on compiling the Estonian wordnet started in 1997 and is still in progress. The work is funded partly by the Estonian Science Foundation and partly within the framework of the Estonian National Programme of Language Technology. The Estonian wordnet currently contains nouns, verbs and some adjectives and proper nouns.

2. WordNet

When designing a semantic database, two fundamental questions have to be faced:

(a) what kind of semantic information is to be stored: linguistic knowledge or world-knowledge?
(b) how is this information stored: as a language-internal system or using a meta-language?

Many semantic databases are not explicit with respect to these distinctions and it is often very difficult to stay consistent. In the case of traditional dictionaries, you might expect to find ‘linguistic knowledge’ telling you what semantic properties determine the exact usage of a word, but a dictionary definition often also contains fragments of word-knowledge. In the case of conceptual networks that try to capture our common-sense knowledge and reasoning, you might expect to only find ‘word-knowledge’ but what is defined are often the same words (Bloksma et al. 1996).

By focusing on historical evidence, the standard dictionaries neglected questions concerning the synchronic organization of lexical knowledge. Beginning with word association studies at the turn of the 20th century and continuing down to the sophisticated experimental tasks of the past twenty years, psycholinguists have discovered many synchronic properties of the mental lexicon that can be exploited in lexicography (Miller et al. 1990).

In 1985 a group of psychologists and linguists at Princeton University undertook to develop a lexical database along lines suggested by these investigations. G. A. Miller and others created WordNet — a semantic database of English. Inasmuch as it instantiates hypotheses based on results of psycholinguistic research, WordNet can be said to be a dictionary based on psycholinguistic principles (Fellbaum 1998).
In WordNet fundamental differences in the semantic organization of syntactic categories can be clearly seen and systematically exploited. Nouns are organized in lexical memory as topical hierarchies, adjectives are organized as N-dimensional hyperspaces, and verbs are organized by a variety of entailment relations. The most ambitious feature of WordNet is its attempt to organize lexical information in terms of word meanings, rather than word forms (Miller et al. 1993).

3. Terms

A word can, and very often does, have several meanings. As WordNet is based on word meaning, all of the words that can express a given sense are grouped together in a SYNONYM SET, alternatively called a SYNSET for short.

For example, when searching for the Estonian word ‘viis’ it is possible to find at least four different meanings, i.e. representatives of four synsets:

- Viis_1_5_1: numeral 5 (the meaning is the number five)
- Viis_2, väga hea_1: kõrgeim hinn eesti koolisüsteemis (the meaning is five – the highest grade in the Estonian school system)
- Viis_3, meloodia_1: helide jõrgnevus muusikas (the meaning is melody – a sequence of sounds in music)
- Viis_4, mood_1, stil_1, laad_1: väljakujunenud, iseloomulik või äratuntav tegutsemislaad või olemisviis (the meaning is manner, style, mode – a specific, known manner of performance)

WordNet is organised according to semantic relations. Since a semantic relation is a relation between meanings and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets.

The database is restricted to the relations suggested by psycholinguistic data: synonyms, antonyms, hyponyms, hyperonyms, meronyms, holonyms, causation, etc. These and other similar relations serve to organise the mental lexicon.

Some examples:

- Hierarchical: koer (dog) HAS_HYPERONYM loom (animal)
- Antonymy: halb (bad) ANTONYM hea (good)
- Part/whole: koer (dog) HAS_MERONYM saba (tail)
- Role/involved: õpetaja (teacher) ROLE õpetama (to teach)
- Causation: tapma (to kill) CAUSE surema (to die)

Just as is the case in the EuroWordNet project, in the Estonian wordnet there are also more than 50 different semantic relations, most of which are reciprocated:

If A HAS_HYPERONYM B, then B HAS_HYponYM A.
4. EuroWordNet – a multilingual semantic database

1996 saw the commencement of EuroWordNet (EWN), a project supported by the European Union and aimed at developing a multilingual database with basic semantic relations between words from several European languages (initially Dutch, Spanish, Italian and English, and from 1998 also Czech, French, German and Estonian). The wordnets are stored in a central lexical database system and the word meanings are linked as in the Princeton WordNet. Each wordnet represents a unique language-internal system of lexicalisations (Vossen 1998b).

The design of the EuroWordNet database is based primarily on the structure of the Princeton WordNet, specifically version WordNet1.5. The notion of a synset and the main semantic relations have been taken over in EuroWordNet. However, some specific changes have been made to the design of the database, which are mainly motivated by the following objectives:

1) to create a multilingual database;
2) to maintain language-specific relations in the wordnets;
3) to achieve maximal compatibility across the different resources;
4) to build the wordnets relatively independently (re-)using existing resources.

To be able to maintain the language-specific structures and to allow for the separate development of independent resources, a distinction is made between the language-specific modules and a separate language-independent module.

Each language-specific wordnet is structured along the same lines; WordNet synonyms are grouped in synsets, which in turn are related by means of basic semantic relations such as hyponymy (between specific and more general concepts) and meronymy (between parts and wholes). By means of these relations all meanings can be interconnected, constituting a huge network or wordnet.

Each language module represents an autonomous and unique language-specific system of language-internal relations between synsets. Nevertheless, some specific measures have been taken to improve the compatibility of the different language-specific wordnets, such as the definition of a common set of so-called Base Concepts (BC). Base Concepts are meanings (i.e. synsets) that play a major role in the wordnets (Vossen et al. 1997).

Equivalence relations between the synsets in different languages and WordNet1.5 are set out in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual wordnets has at least one equivalence relation with a record in the ILI, either directly or indirectly via other related synsets. Language-specific synsets linked to the same ILI record should thus be equivalent across the languages. Synsets linked to the same WordNet1.5 synset should be equivalent or close in meaning and can then be compared.
5. Estonian Wordnet

Compilation of the Estonian wordnet (EstWN) began in 1997 and the work is ongoing. It is important to point out that other partners in the EWN project have created their wordnets automatically whereas we have preferred manual work. One reason for this is the absence of suitable electronic material, while there is also the fact that the results produced this way are more accurate.

In addition to general vocabulary, legal vocabulary is also presented in order to facilitate the precise translation of legal texts. Table 1 gives an overview of the current contents of the Estonian wordnet.

Table 1: Number of synsets per different part-of-speech in EstWN

<table>
<thead>
<tr>
<th>SYNSETS</th>
<th>Proper Nouns</th>
<th>Nouns</th>
<th>Verbs</th>
<th>Adjectives</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meanings (variants)</td>
<td>446</td>
<td>6734</td>
<td>2757</td>
<td>307</td>
<td>10244</td>
</tr>
<tr>
<td>Meanings per synset</td>
<td>1.06</td>
<td>1.67</td>
<td>2.08</td>
<td>1.69</td>
<td>1.75</td>
</tr>
<tr>
<td>Different words and compounds</td>
<td>470</td>
<td>9518</td>
<td>3797</td>
<td>419</td>
<td>14204</td>
</tr>
<tr>
<td>Meanings per word</td>
<td>1</td>
<td>1.18</td>
<td>1.51</td>
<td>1.24</td>
<td>1.26</td>
</tr>
<tr>
<td>Semantic relations</td>
<td>474</td>
<td>15024</td>
<td>5596</td>
<td>538</td>
<td>21632</td>
</tr>
<tr>
<td>Semantic relations per synset</td>
<td>1.06</td>
<td>2.23</td>
<td>2.03</td>
<td>1.75</td>
<td>2.11</td>
</tr>
</tbody>
</table>

The major lexical information which is necessary for compiling thesauri comes from monolingual explanatory and/or sense-distinguishing dictionaries or synonym dictionaries. However, there is only a very restricted number of these available in Estonian. Machine-readable dictionaries are also difficult to come by.

Data for the compilation of the Estonian WordNet is obtained from the following sources:

(a) to find word meanings, explanations and examples, the Estonian Explanatory Dictionary, which is unfortunately not a completely machine-readable dictionary, is used
(b) in the case of synonymy and antonymy relations, synonym and antonym dictionaries are used
(c) to link concepts with the Inter Lingual Index, bilingual dictionaries are used
(d) other Estonian language technology tools are used. For example, word frequency records are compiled on the basis of the Corpus of Estonian Literary Language (CELL, 1 million words). Information from the Corpus is also used to define the different meanings of a word and quotations from the Corpus are used as examples.

6. Compilation of the Estonian Wordnet

In general, the wordnet was built in two major cycles as indicated in Figure 1. Each cycle consisted of a building phase and a comparison phase. When creating the database itself, we used the EWN Database Editor Polaris (Louw 1998).
6.1. Words from texts

The initial version of EstWN is intended to cover the Estonian base vocabulary. We refrained from translating the Princeton WordNet, with the exception of the Base Concepts, with the main emphasis being placed on frequency lists of word forms. The base vocabulary was determined by statistical analysis of the Corpus of Estonian Literary Language.

Lexical entries in EstWN are presented in nominal singular form for nouns and supine form for verbs. In real texts, the words appear in a wide range of forms. For example, there are fourteen cases in Estonian, yielding 28 forms in every noun paradigm. On average, 45% of the word forms in Estonian texts are morphologically ambiguous and this fact produces irrelevant data in the lexicon word list.

The number of compounds in Estonian is indefinite. It is quite easy for a user to make up compounds that are not found in any dictionary but are still understandable to other users. It is not practical for all imaginable compounds to be entered in wordnet. The choice can be made by analysing the components of a compound: if the components are of low frequency as independent words in corpora, then the compounds they form should be added into the lexicon.

Estonian is a flective language with a free word order which complicates the process of identifying phrases from corpus texts. The elements of a phrase can be scattered around the
sentence in an unpredictable order. Multi-word expressions are included in EstWN if they make up a conceptual unit and are commonly used as lexical units.

6.2. A set of synonymous word meanings

One of the problems encountered concerned dividing the meanings of a word into different synsets in EstWN. Due to the lexicographical sources used, over-differentiation (word meaning marked as too specific) and over-generalisation (word meaning marked as too general) may occur. Traditional dictionary entries are centred on the word itself and all syntactically or pragmatically specific meanings are represented, but semantically they are not always significant.

Which new words or meanings should be concentrated on when EstWN is upgraded? It is essential that words which are actually used in texts are added.

It became clear that the results of word sense disambiguation (WSD) of corpus texts were well-suited in adding missing synsets and senses to our wordnet. There were significant inconsistencies in the opinions of the people who disambiguated the texts. This identifies the most problematic entries in EstWN and the need to reconsider the borders of meaning of some concepts.

6.3. Language-dependent lexicalisation vs Inter-Lingual Index

As previously mentioned, the wordnets of different languages are linked via the Inter Lingual Index in EWN. However, lexicalisation features or the borders of any one concrete word meaning are in many cases different from language to language and it often proves impossible to find an exact and equal relation (eq_synonym). Fortunately, it is possible to use other eq-relations to describe more precisely the difference between source language and ILI synsets.

The Estonian language, as a Finno-Ugric language, is different from most Indo-European languages, such as those in EuroWordNet. Some specific lexicalisation patterns appear when we try to link Estonian synsets to the ILI.

In the case of verbs, there is a fairly regular discrepancy between English and Estonian in the expression of causativity. In Estonian, the causative counterparts of non-causative verbs are regularly lexicalised. A causative suffix added to an intransitive non-causative verb produces a new lexical entry which is syntactically transitive and causative in meaning. In addition, there are cases where English verbs have only causative or only non-causative meaning, whereas in Estonian both meanings are lexicalised and used productively. For instance, the English verb 'to spend (money)' is translated as 'kulutama', but there is also the non-causative 'kuluma' (also about money) for which no exact English lexicalised counterpart exists (the dictionary translation is 'be spent' but the Estonian 'kuluma' has no passive meaning). The same is also true of 'levitama' (to disseminate, distribute) and the non-causative 'levima'. The problem arises from the ILI synsets where there is no causativity/non-causativity marked, making it impossible to use either eq cau ses or eq_is caused_by relations for linking source language and ILI synsets.
For example, the Estonian equivalents of ‘to move’ are ‘liikuma’ and ‘liigutama’. The second of these is causative. In Estonian, these differences are systematically lexicalized in a different manner to that of other languages (Vider 2002).

Also, onomatopoetic-descriptive words are more detailed in Estonian. One example: in Estonian, a bear ‘mõmiseb’ and a lion ‘mõirgab’. In English they both ‘roar’.

7. Conclusion

The Estonian wordnet can be used, amongst other things, for the retrieval of monolingual and cross-lingual information which can be searched by meaning and concept. A wordnet-type thesaurus is also helpful as a translation tool or for compiling other, different kind of lexicons. It is very important for Estonian language technology to have such a lexical-semantic resource.

This paper has pointed out some of the difficulties which have appeared during the practical compilation of an Estonian wordnet-type thesaurus.

References


