Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet

Heili Orav and Kadri Vider

University of Tartu, Department of General Linguistics,
Liivi 2, 50409 Tartu, Estonia

Abstract. One source of Estonian WordNet have been corpora of Estonian. On the other hand, we get interested in word sense disambiguation, and about 100,000 words in corpora are manually disambiguated according to Estonian WordNet senses. The aim of this paper is to explain some theoretical problems that "do not work well in practice". These include the differentiation of word senses, metaphors, and conceptual word combinations.

1 Introduction

By now the research group of computational linguistics at the University of Tartu has worked six years on the thesaurus of Standard Estonian or the Estonian WordNet (EstWN)[1]

Although the thesaurus covers only about ten thousand concepts, experiments in the disambiguation of textual words show that thesaurus entries cover the majority of senses of Estonian core vocabulary [1].

When setting up the Estonian WordNet we followed the principles of Princeton WordNet and EuroWordnet. For a more detailed discussion see Kahusk and Vider [2].

The existing thesaurus was used as the Estonian basic lexicon for SENSEVAL-2 contest[2].

The aim of this paper is to point out some theoretical problems that 'do not work well in practice'. These include the differentiation of word senses, metaphors, and conceptual word combinations.

2 Estonian WordNet and Word Sense Disambiguation Task

Lexically the thesaurus is derived from the existing traditional dictionaries (mainly the "Explanatory Dictionary of Estonian") and a text corpus (providing information about usage).

At present the Estonian WordNet contains about ten thousand synsets: mostly noun (66 %) and verb concepts (27 %), but also a limited number of adjectives (2.6 %) and proper nouns (4.4 %). Each synset has more than two semantic relations; hyponymic and hyperonymic relations predominate.

[1] This paper is based on work supported in part by the Estonian Science Foundation under Grant 5534 and by Estonian State Target Financing R&D project number 0182541s03 "Eesti keele arvutimudelid ja keeleressursid: teoreetilised ja rakenduslikud aspektid."

[2] See http://www.sle.sharp.co.uk/senseval2/

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 285–290.

We got interested in word sense disambiguation (WSD) couple of years ago and at present time we have a corpus of about 100,000 manually disambiguated textual words. The texts were taken from the Corpus of Estonian Literary Language. The sense numbers of the Estonian thesaurus were used to disambiguate only nouns and verbs because the including of adjectives in the thesaurus began only recently.

At present we are adding new word senses to the EstWN on the basis of word sense disambiguation. These findings reveal some theoretical and practical drawbacks in setting up the thesaurus.

3 Too Broad or Too Narrow?

When looking up the meaning of a specific textual word in the thesaurus, it often seems that the meaning recorded in the thesaurus is either too specific or too general for the given context. The disambiguation of word senses in a text reveals quite clearly that a broader or narrower meaning of the word is synonymous with the senses of other words in a concrete usage but not in the conceptual system.

Let us take a look at the example sentence

Example 1. Laps läks kooli ‘the child went to school’,

where it is irrelevant whether the child went to school as an educational institution or a building, or actually both were meant. At the same time the sentence

Example 2. Linn on ehitanud sel aastal juba kolm kooli ‘this year the town has already built three schools’

means that in this case only the school building is meant.

kool_1 [polysemic sense that applies both to the institution and the building]
⇒**kool_2** [school building]
⇒**kool_3** [educational institution]

Fig. 1. Hyponymic senses for *kool* (‘school’)

If the thesaurus provides the hyponymic and hyperonymic senses for the word *kool* ‘school’ (see Figure 1), there will be more than enough different senses of *kool*. The second and the third senses (narrower senses) are covered by sense 1 as a more general one. In the case of manual disambiguation the marking of the more general sense (sense 1) is usually justified. Sense 2 will be needed only for such cases as example sentence 2. However, if the synset including sense 1 has both the building and institution as its hyperonyms, then *kool* in sentence 2 could be disambiguated correctly by means of sense 1 as well.

In a semantically related thesaurus like WordNet each synset can have only a single hyperonymic relation. Therefore, it is highly inconvenient to present regular polysemy, and one tries to avoid polysemy by adding broader or narrower senses of the same word. This, however, creates for the semantic disambiguator a disturbingly large number of senses that are rarely used and are difficult to distinguish from one another.

One way to decide whether the addition of a narrower or broader sense to the thesaurus is justified is to find translation equivalents for the meanings of textual words. For example, the Estonian verb *kuduma* has at least two clearly distinguishable senses that belong into different synsets in English Wordnet (see Figure 2).

kuduma_1 *weave, tissue* [of textiles; create a piece of cloth by interlacing strands, such as wool or cotton]
kuduma_2 *knit* [make textiles by knitting]

Fig. 2. Different senses of verb *kuduma* belong into different synsets, and have different literals in English ('weave' and 'knit')

The above-mentioned WordNet senses correspond to subdivisions 1.a. and 1.b. of entry *kuduma* in "Explanatory Dictionary of Estonian". It means that they are regarded as rather specific subsenses of the more general meaning of *kuduma* 1. in Estonian. However, it is difficult to find an example of the verb *kuduma* in the text, where it is not important whether one is weaving a fabric or knitting using knitting needles. It shows that the thesaurus has to introduce two clearly distinguishable senses of *kuduma* (in addition to senses 2 and 3 provided in the explanatory dictionary). For the same reason, one might omit the more general sense of *kuduma* (sense 1 in the explanatory dictionary).

Naturally it is difficult and perhaps even impossible to distinguish the meanings in one language from the perspective of many other languages, and there is no good reason for preferring a certain language for translation equivalents for the purpose of a monolingual thesaurus. However, one should consider the use of translation equivalents as a possibility if the thesaurus makers disagree on whether the senses in the thesaurus are too narrow or too broad.

4 What Should We Do with Metaphors?

Metaphors and metaphorical meanings of words are a topical issue in linguistics and lexicology. Even the well-known psycholinguist and founder of WordNet George A. Miller provided a thorough classification of metaphors in "Metaphor and Thought" [3].

Metaphors present an appropriate touchstone for a thesaurus. They raise the question whether the senses arising from the metaphorical use of words should be added as new meanings to the thesaurus or not. Their occurrence in text is really rather unpredictable and chaotic. And if we add the metaphorical uses to the thesaurus, then how should we explain them properly. As is known, the understanding of a metaphor depends on the context.

Below you will find an example from our semantically disambiguated corpus:

Example 3.

> *Loopis taas oma murruvahus latvu vastu kaldakivisid, peksis neid vanu vaenlasi, kes kuidagi ei tahtnud endid veerevate lainemägede teelt ära koristada. (tkt0034)*
>
> 'it was once again hurling its foamy tops against the rocks, it was lashing its old enemies who wouldn't make way to the rolling mountainous waves'

The author has described a stormy sea. In the case of manual semantic disambiguation one would ask the question what do the words *latv* 'treetop', *vaenlane* 'enemy', *loopima* 'hurl', *peksma* 'beat, lash', *koristama* 'clean, clear' mean. One might presume that these words have specific meanings in the thesaurus that cannot be extended to the textual meanings without pointing out their metaphoricalness.

It is possible to distinguish between two main types of knowledge in the comprehension of a text [4]:

1. semantic knowledge is knowledge of extralinguistic reality;
2. pragmatic knowledge is knowledge regulating communication (social norms, conventions).

Because EstWN is based on the existing traditional dictionaries and a text corpus (providing usage information), one might suppose that the semantic information in the database reflects semantic knowledge.

The addition of metaphors to the thesaurus would make it a thesaurus that combines semantic and pragmatic combinations. It would increase the size of the thesaurus to a remarkable degree. For this reason until now we have tried to avoid the addition of metaphors, but problems are opened.

5 Conceptual Word Combinations

Conceptual word combinations present another problem in the disambiguation of word senses. The thesaurus includes 984 such combinations as entries, three quarters of them being phrasal and phraseological verbs. They are mostly two-word combinations, but there are also some three- and even four-word combinations as well.

Comparison with the database of Estonian collocations (multi-word units, see Kaalep & Muischnek [5])[3] shows that 635 items overlap as phrasal and phrsaeological verbs and only six as noun expressions.

Why do we call them conceptual word combinations and not phraseological units? Phraseology proceeds from language use, and a phraseological unit is a combination that is always used together but the meaning of which differs from the sum of the meanings of its constituents [6]. A large number of conceptual word combinations in the thesaurus are phraseological units as well (metaphorical phraseological verbs, for example). On the other hand, the thesaurus entries include many combinations constituting a conceptual whole. They cannot be regarded as phraseological units because their meaning arises from the meaning of their constituents, and they are not collocations in statistcal terms.

Conceptual word combinations became thesaurus entries as:

[3] See `http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html`.

1. synonyms (e.g. *meenutama, meelde tuletama* 'recall, remember'; *üllitama, välja andma* 'publish');
2. specific hierarchical nodes (e.g. *emotsionaalne seisund* 'emotional state', *ruumiline omadus* 'spatial characteristic', *üleloomulik olend* 'supernatural creature', *suuruse või koguse muutmine* 'modification of size or amount';
3. technical terms (e.g. *ilmaütlev kääne* 'abessive case', *damaskuse teras* 'Damascus steel', *kreeka tähestik* 'Greek alphabet';
4. explanations (e.g. *kultiveerima, kultuurina kasvatama* 'cultivate, grow as a culture', *naer, naeru hääl* 'laughter, sound of laughter',
   hääletaja, pöidlaküüdiga sõitja 'hitchhiker, a person thumbing a lift'.

Synonyms (1) and technical terms (3) are the only groups of word combinations that justify their inclusion in the thesaurus from the perspective of word sense disambiguation. From the same perspective one can only welcome the fact that two thirds of the word combinations included in the thesaurus can be also found in the database of multi-word units. The latter database is likely to serve in the future as a basis for morphological and syntactic recognition of word combinations in texts. Once the computational analysis of previous levels is able to recognize multi-word units in a text, it will be possible to find the matching senses in the thesaurus. Because it is likely that the components of noun combinations occur close to each other in a text, formally it is easier to spot them first automatically and then compare them against the word list of the thesaurus. The recognition of verb combinations, however, is still an unmanageable task for lemmatizers. Due to inadequate pre-processing at the present level of semantic disambiguation the conceptual word combinations are provided with wrong meanings both in the course of automatic tagging and sometimes also in manual tagging. On the other hand, the thesaurus includes as synonyms a certain number of (verb) combinations that are not included in the database of multi-word units because of their rare occurrence. However, these combinations are essential for the thesaurus (e.g. *arvamusele jõudma* 'reach an opinion', *keelele tulema* 'come on the tip of one's tongue', *ühel meelel olema* 'be of the same opinion'). Thus, these combinations should be included in the database of multi-word units in cooperation with the creators of this database.

Thus, the combinations in groups (2) and (4) seem useless from the perspective of word sense disambiguation. If we define these groups on the basis of absence from the database of multi-word units, then it will be easy to find a good reason for carrying out a semantic analysis by components once the fixed combination recognition software is complete. There is strong likelihood that this is going to happen to the explanatory conceptual combinations of group (d). In addition, one should also consider their suitability as thesaurus entries. It would be reasonable to place such combinations in the explanation field of a synonymous entry.

6 Conclusions

It appears that the creation of a concept-based thesaurus is not as easy as it seems at first sight. The main problems in setting up a thesaurus include:

- under- or over-differentiation;
- metaphors;
- conceptual word combinations.

The practical use of the thesaurus in WSD task showed that the senses based on the traditional defining dictionary and the intuition of lexicographers may be either too narrow or too broad. This fact compels the thesaurus makers to order the word senses both in the thesaurus and to think about the reliability of the previous theoretical views.

At the same time semantic disambiguation experiments show that the meaning of the sentence and the meaning of the lexical words constituting the sentence are largely dependent on the functional words. Unfortunately, the latter are not included in the thesaurus, and the semantic tagging system that is based on the thesaurus does not take them into account. Prior recognition of conceptual word combinations would make at least one part of such meaning-differentiating units 'visible' for word sense disambiguation.

References

1. Kahusk, N., Orav, H., Õim, H.: Sensiting inflectionality: Estonian task for SENSEVAL-2. In: Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguating Systems, Toulouse, France, CNRS—Institut de Recerche en Informatique de Toulouse, and Univeresité des Sciences Sociales (2001) 25–28.
2. Kahusk, N., Vider, K. In: Estonian WordNet Benefits from Word Sense Disambiguation. Central Institute of Indian Languages, Mysore, India (2002) 26–31.
3. Miller, G. A. In: Images and models, similes and metaphors. 2nd edn. Cambridge University Press (1979).
4. Õim, H.: Семантика и теория понимания языка. Анализ лексики и текстов директивного общения эстонского языка. PhD thesis, University of Tartu (1983).
5. Kaalep, H.J., Muischnek, K. In: Inconsistent Selectional Criteria in Semi-automatic Multi-word Unit Extraction. Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest (2003) 27–36.
6. Õim, A.: Fraseoloogiasõnaraamat. Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut, Tallinn, Estonia (1993).