Some Problems in Estonian Wordnet

Kadri Vider
kadriv@madli.ut.ee
Department of General Linguistics
University of Tartu

WN is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. Nouns, verbs, adjectives and adverbs are organised into synonym sets, each representing one underlying lexical concept. Different kinds of semantic relations link the synonym sets (synsets). WN is based on **word meaning**; all of the words that can express a given sense are grouped together in a SYNONYM SET, or SYNSET (see also Vider K., Orav H. 1996. WORDNET: An On-line Lexical Database. *Papers of the First Swiss-Estonian Student Workshop on Computational and Theoretical Linguistics*: 64-68).

English WN is made by the psycholinguists of the Princeton University. Today it is of wide interest also to the linguists of other languages. The EuroWordNet project is currently producing a generic multilingual semantic database, which is the first in its kind. It contains the basic semantic information for Dutch, Italian, Spanish and English, while each of these resources is linked to a shared inter-lingua. EuroWordNet-2 extend the project with a French and German wordnet. and with two Eastern-Middle European sites - Estonian and Czech - that are involved producing wordnets for their national languages. Estonian WordNet will join the project EuroWordNet-2 as the builder from the beginning of January 1998. In the framework of the project of Estonian language technology the Estonian WordNet has to be created during the years 1997-2000. It will incorporate in addition to the general vocabulary also legal vocabulary as to facilitate the precise translation of legal texts.

Unfortunately the compilation of a dictionary is not as easy as it seems in theory. Already when experimenting with the first twenty or so words a number of problems cropped up that can generally be divided into three groups:

1. Problems with sources
2. Problems with synsets

3. Problems with semantic relations

1. Problems with sources

The data for the compilation of the Estonian WordNet are got from the following sources:

a) word frequency records are compiled on the basis of the Corpus of Written Estonian (which contains one million words); the materials of the Corpus are also used to define the different meanings of a word and quotations from the Corpus are used as examples;
b) in case of synonymy and antonymy relations the dictionaries of synonyms and antonyms are used;
c) to get the word meanings, explanations and examples the Estonian Explanatory/Monolingual Dictionary, which is unfortunately not a completely machine readable dictionary, is used.

1.1. Frequency records (lists)

As it is with all the dictionaries, similarly the compilation of the Estonian WordNet begins with the putting together of the word lists. In a thesaurus like WN mainly substantives, verbs, adjectives and adverbs are dealt with. Thus pronouns, conjunctions and the other helpwords that are on the top of the frequency list are left out of the semantic consideration, neither have they been included in the word list of the thesaurus. For the compilation of the thesaurus word list the absolute occurrences in the corpus are in fact not as necessary as the probable evaluation of the ranking of a word in the frequency list of a corresponding wordgroup.

Our aim is to present one thousand of so-called base concepts in the WN format by the end of this year. These base concepts share the features of having high frequencies, poor definitions, a high degree of polysemy, a high number of hyponyms appearing in the higher levels of the taxonomy. Some issues of EuroWordNet define that main criterion of extracting a base concept will be its frequency as definition word and corpus-frequency. It is impossible to count the appearance of a word as a definition word in Estonian, because of the lack of the complete electronic issue of the explanatory dictionary, whereas neither in the existent part of it (beginning with K) the definitions are not consistently tagged.

Moving further to the other criterion - what are the high-frequency words? It is easy to get the frequency list of the occurring forms from the corpus texts, good means exist also for the morphological analysis of the forms. But especially among the more frequent forms there are many such forms that can be analysed morphologically in various ways.

(1)

(450.000 sample, verb forms F>10

output of Estonian morphological analyser,

A – adjective, D – adverb, P – pronoun, S – noun, V - verb)

```
1986_OMA    oma //_A_ sg g, sg n, sg p, //    oma //_D_ //
oma //_P_ //    omama //_V_ o, //
26_OMAD    oma //_A_ pl n, //    omama //_V_ d, //
24_OMAKS    oma //_A_ sg tr, //    omama //_V_ ks, //
```

Thus the mechanical adding together of ambiguous forms can sometimes lead to unreliable results, in the given example the occurrence frequency of the verb 'omama' (own, possess) has been raised by the form 'oma' (which can occur in Estonian as noun and as adverb as well). While the next forms in order of frequency of the verb 'omama' have relatively smaller occurrence frequency, it is obvious even without manual controlling that the verb 'omama' does not belong among the MOST frequent verbs. (This cannot be said with certainty about the concept 'omama' (possess, own)). The second example is a bit more difficult:

(2)

(450.000 sample, verb forms F>10

output of Estonian morphological analyser,

A – adjective, D – adverb, P – pronoun, S – noun, V - verb)

```
125_AJA    aeg //_S_ sg g, //    ajama //_V_ o, //
56_AJAKS    aeg //_S_ sg tr, //    ajama //_V_ ks, //
42_AJAS    aeg //_S_ sg in, //    ajama //_V_ s, //
25_AJADA    ajama //_V_ da, //
15_AJAB    ajama //_V_ b, //
14_AJAD    aeg //_S_ pl n, //    ajama //_V_ d, //
10_AJANUD    ajanud //_A_ sg n, //    ajanu //_A_ pl n, //
ajanu //_S_ pl n, //    ajama //_V_ nud, //
10_AJAMA    ajama //_V_ ma, //
10_AETUD    aetud //_A_ sg n, //    aetu //_A_ pl n, //
aetu //_S_ pl n, //    ajama //_V_ tud, //
```

As we can see, the paradigms of the substantive 'aeg' (time) and the verb 'ajama' are mixed up here (verb 'ajama' is very ambiguous, it translates in English as to drive, to incite, to prompt, to stimulate; to impel). In this case we are left with the possibility to check the occurrence of the corresponding forms in the corpus text. This kind of manual work with word forms indicates once more the indispensability of a proper disambiguator.

As our aim in the future is the uniting of Estonian wordnet with WN1.5, we hoped to find from there the list of base concepts. In WN 1.5 the nouns as well as the verbs are divided into logical categories and we decided to stick to the same categories for the time being.

(3)

TOP	WN1.5
#n#tegu#	noun.act
#n#loom#	noun.animal
#n#ese#	noun.artifact
#n#omadus#	noun.attribute
#n#keha#	noun.body
#n#tunnetus#	noun.cognition
#n#suhtlus#	noun.communic
#n#sündmus#	noun.event
#n#tunne#	noun.feeling
#n#toit#	noun.food
#n#rühm#	noun.group
#n#koht#	noun.location
#n#siht#	noun.motive
#n#objekt#	noun.object
#n#isik#	noun.person
#n#nähtus#	noun.phenomenon
#n#taim#	noun.plant
#n#omamine#	noun.possession
#n#protsess#	noun.process
#n#määr#	noun.quantity
#n#suhe#	noun.relation
#n#kuju#	noun.shape
#n#olek#	noun.state
#n#aine#	noun.substance
#n#aeg#	noun.time
#n#keha#	verb.body

#n#muutus#	verb.change
#n#tunnetus#	verb.cognition
#n#suhtlus#	verb.communication
#n#võistlus#	verb.competition
#n#toit#	verb.consumption
#n#kontakt#	verb.contact
#n#looming#	verb.creation
#n#tunne#	verb.emotion
#n#liikumine#	verb.motion
#n#taju#	verb.perception
#n#omamine#	verb.possession
#n#ühiskond#	verb.social
#n#seisund#	verb.stative
#n#ilm#	verb.weather

I tried to find from the WN1.5 *.dat-files the kind of synsets that are lacking superordinate terms, consequently therefore they themselves must be on the highest position in the hierarchy. The results were somewhat surprising: nouns had 11 and verbs had 339 synsets of this kind. Evidently it can be explained with the help of the WN compilers' claim, that nouns are organised in lexical memory as topical hierarchies, verbs are organised by a variety of entailment relations and adjectives are organised as N-dimensional hyperspaces.

(4)

Noun top synsets

entity - something having concrete existence; living or nonliving

psychological feature - a feature of the mental life of a living organism

abstraction - a concept formed by extracting common features from examples

location - a point or extent in space

shape, **form** - the spatial arrangement of something as distinct from its substance

state - the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state"

event - something that happens at a given place and time

act, **human action, human activity** - something that people do or cause to happen

group, grouping - any number of entities (members) considered as a unit

possession - anything owned or possessed

phenomenon - any state or process known through the senses rather than by intuition or reasoning

1.2.Multi-word expressions

If we want a reliable evaluation about the occurrence of one or the other frequent word in the context of the corpus, we should take into account the fact that Estonian language contains plenty of compound and expression verbs as well as idiomatic multiword compounds, that cannot be left without attention, for they have different meanings than their headword and thus belong to different synsets. While Estonian language lacks strict wordorder and while in certain forms the compound verbs are spelled together and in others separately, even the automatic enumeration of a word's collocations is practically of no use.

(5)

ANDMA

#andeks_andma,andestama# - forgive, pardon

ILU\stkt “Ma annan talle andeks!"

ILU\stkt Sulle on su inetud teod andeks antud," kuulutas Sirje.

ILU\stkt Ja pattude andeksandmise ja õndsuse...

ILU\stkt Anna siis oma poisikesele andeks!

#välja_andma,üllitama# - give out, issue

AJA\stat Teisel päeval andis iga koondrühm välja välklehe.

At the same time it is not always possible to convey the most general notions by means of one word. In the corpus texts that show the USAGE of the language, the most general notions need not be the most frequent ones.

(6)

OLEMA 18416

be,occupy_a_certain_position,occupy_a_certain_area

equal, be_identical_to, be

have, have_got, hold
own, have, possess, have_possession_of
be, work
exist, be
be, have_the_quality_of_being
be, occur
originate_in, come_from, hail_from, be_from

aktiivne_olema, ametis_olema, ärkvel_olema, ilma_olema, kindel_olema, kogenud_olema, nõus_olema, olemas_olema, omanik_olema, paigal_olema, palgal_olema, parem_olema, pärit_olema, peidus_olema, pime_olema, raevunud_olema, rahulik_olema, sama_olema, seotud_olema, tuttav_olema, ühenduses_olema, üldkehtiv_olema, ülekaalus_olema, õpilane_olema, valvas_olema, veendunud_olema, võrdne_olema

VÕIMA 3129

OMAMA 2036 have, have_got, hold
own, have, possess, have_possession_of

SAAMA 2034 become

aru_saama, kasu_saama, kuulda_saama, lahti_saama, teatavaks_saama, tugevaks_saama

PIDAMA 1706 consider, count, weigh
observe, celebrate, keep

kinni_pidama, kirjavahetust_pidama, meeles_pidama, paremaks_pidama, ülal_pidama, vastu_pidama

TULEMA 1629 come, come_up
arrive, get, come

ette_tulema, kokku_tulema, nähtavale_tulema, sisse_tulema, välja_tulema

TEGEMA 1559 make, create

edusamme_tegema, häält_tegema, häbi_tegema, heameelt_tegema, keeruliseks_tegema, kingitust_tegema, lahti_tegema, muret_tegema, nähtavaks_tegema, olematuks_tegema, sõjakäiku_tegema, tööd_tegema, tundlikuks_tegema, ümber_tegema, vahet_tegema, valesti_tegema, vigu_tegema

MINEMA 929 go, go_away, depart, travel_away move, go

kaotsi_minema, kaubaks_minema, lahku_minema, magama_minema, sõtta_minema, voodisse_minema

JÄÄMA 862 have, have_left
persist, remain, stay

alla_jääma, ellu_jääma, ilma_jääma, kindlaks_jääma, maha_jääma, nõrgaks_jääma, ootama_jääma, paigale_jääma

ANDMA 859 give, cause_to_have

alla_andma, ära_andma, eetrisse_andma, järele_andma, maitset_andma, nime_andma, õnnistust_andma, puhkust_andma, tööd_andma, välja_andma, värvi_andma

We translated the 339 top-synsets of the verbs that we found, into Estonian on the principle, that we tried to find the equivalent Estonian words to the meanings expressed by the synsets (we didn't translate the members of the synset one by one!). In the current example (6) the frequency list of the verbs from corpus is given (it is composed by counting ambiguous forms!), after that are given the top-synsets that had in their Estonian equivalent the corresponding verb alone. Further below the corresponding verb as a part of the compound verb in translation have been given (the occurrence of those compound verbs in the corpus is not yet worked through).

Briefly, we can say that it is possible to get from the corpus and from other this kind of collections of data the more frequently occurring expressions which in the hierarchy of concepts form no more than the base level of usage. The base concepts belonging to the top of the superordination/subordination hierarchies can be found out only after thorough study of the frequent words.

3. Problems with synsets

The design of the EWN-database is first of all based on the structure of the Princeton WN1.5. A line in the data file mainly consists of three components:

<identifier> <synset content> <semantic relations>

Princeton WN line delimiters are single spaces, because the information in it is on a single level. There are several levels in EuroWN input/output format, but most of them consist only different kind of lexical labels like morphological features and usage. Estonian WordNet format is on the middle of those two polarity, we use three types of delimiters today (# ; ,).

(7)

identifier			00000042
#			#
part of speech (n, v, a, d, i)			v
#			#
S		word1	hakkama5
Y		,	,
N		multi_word_expression2	peale_hakkama
S		,	,
E		word4	algama1
T		,	,
		...	pihta_hakkama
C	;		;
O		"example"	"Kool on juba alanud."
N		,	
T		"example"	
E		,	
N		...	
T	;		;
	'explanation'		'algust SAAMA'
#			#
semantic relation			TOP00000038
#			#
semantic relation			ANT00000143
#			#
semantic relation			HYP00000144
#			#
...			

Example:
00000142#v#hakkama5,algama1,peale_hakkama,pihta_hakkama;"Kool on juba alanud.";'algust SAAMA.'#TOP0000038#ANT00000143# HYP00000144

Synsets, however, are made word by word. A word is taken (e.g.) and checked in which lines (=synsets) it already occurs, if some of them satisfies the sense selected from the corpus or the explanatory dictionary, the word is given a sense code in this synset; if none of the existent lines is satisfactory a new synset is created.

Through the synonymy relations the number of the words that were chosen initially will increase by the addition of an indefinite amount of new words, that in their turn have to be taken under scrutiny and decided upon which of them are to be divided into senses.

It means that n words are distributed by their senses into synsets and that there are definitely more senses than there are words.

3. Problems with semantic relations

The relations do not rely on any specific knowledge-representation formalism and are expected to form the backbone of any knowledge system of the future. We have taken as our aim to be directed in the selection of semantic relations by the choices of the EuroWordNet. The problems on that plane have to do with the decisions upon what kind of semantic relations should we consider as important and how should we determine the relation between two synsets. The instructions of EuroWordNet offer a considerably bigger number of semantic relations as compared to the Princeton WN.

(8)

Relation type	Abbreviation
NEAR_SYNONYM	NSN
XPOS_NEAR_SYNONYM	XSN
HAS_HYPONYM	HYP
HAS_HYPERONYM	HPR
HAS_XPOS_HYPONYM	XYP
HAS_XPOS_HYPERONYM	XPR
HAS_MERONYM	MER
HAS_HOLONYM	HOL
HAS_MERO_PART	MPA
HAS_MERO_MEMBER	MME
HAS_MERO_POSITION	MPO
HAS_MERO_MADEOF	MMA

HAS_MERO_LOCATION	MLO
HAS_HOLO_PART	HPA
HAS_HOLO_MEMBER	HME
HAS_HOLO_POSITION	HPO
HAS_HOLO_MADEOF	HMA
HAS_HOLO_LOCATION	HLO
ANTONYM	ANT
NEAR_ANTONYM	NAN
XPOS_NEAR_ANTONYM	XAN
IS_CAUSED_BY	CAB
CAUSES	CAU
IS_SUBEVENT_OF	SEO
HAS_SUBEVENT	SEV
INVOLVED	INV
ROLE	ROL
INVOLVED_AGENT	IAG
INVOLVED_INSTRUMENT	IIN
INVOLVED_PATIENT	IPA
INVOLVED_LOCATION	ILO
INVOLVED_DIRECTION	IDI
INVOLVED_SOURCE_DIRECTION	ISD
INVOLVED_TARGET_DIRECTION	ITD
ROLE_AGENT	RAG
ROLE_INSTRUMENT	RIN
ROLE_PATIENT	RPA
ROLE_LOCATION	RLO
ROLE_DIRECTION	RDI
ROLE_SOURCE_DIRECTION	RSD
ROLE_TARGET_DIRECTION	RTD
STATE_OF	STO
BE_IN_STATE	BIS
HAS_DERIVED	DER
IS_DERIVED_FROM	DEF
FUZZYNYM	FZZ
XPOS_FUZZYNYM	XFZ
HAS_INSTANCE	INS
BELONGS_TO_CLASS	BTC

The relations over the wordgroup borders have been involved (this brings about a big number of synsets that are essentially doubling each other e.g. in Estonian we can derive from the verbs nouns that are same in meaning with the help of the suffix '-mine'.) The relations of superordination/subordination seem to be the most important ones though, especially in the case of legal vocabulary, where a cognisable system of terms has yet to be created. If we had good definitions for every sense, it would be easy to involve such interesting relations like ROLE/INVOLVED, that were missing in the Princeton WN. The workgroup of EuroWordNet has worked out some tests to determine the relations between two synsets.

(9)

Verb test		
Comment:		Hyperonymy/hyponymy between verb synsets
Score		**Test sentence**
yes	a	X is Y+ AdvP/AdjP/NP/PP
no	b	Y is X + AdvP/AdjP/NP/PP
Conditions:		- X is a verb in the infinitive form - Y is a verb in the infinitive form - there is at least one specifying AdvP, NP or PP that applies to the Y-phrase
Example:	a	to run is to go fast
	b	* to go is to run fast
Effect:		{to run} (X) HAS_HYPERONYM {to go} (Y) {to go} (Y) HAS_HYPONYM {to run} (X)

References

Alonge, A. 1996. Definition of the links and subsets for verbs. Final version 6. *EuroWordNet.*

Beckwith, R., Fellbaum, C., Gross, D., Miller, G. 1990. 'WordNet: A Lexical Database Organized on Psycholinguistic Principles.' In Zernik, U. (Ed.), *Using On-line Resources to Build a Lexicon*. Chapter 9, 211-231, Hillsdale, NJ: Erlbaum.

Bloksma, L., Diez-Orzas, P. L., Vossen, P. 1996. User Requirements and Functional Specification of the EuroWordNet project.

Diez-Orzas, P. L., Forest, P., Louw, M. 1996. High-level Architecture of the EuroWordNet Database. A Novell ConceptNet-based semantic network. Final version 7. *EuroWordNet.*

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J. 1990. Introduction to WordNet: An On-line Lexical database. *International Journal of Lexicography*, 3: 235-312.

Miller, G., Fellbaum, C. 1991, Semantic networks of English. *Cognition*, 41: 197-229

Vider K., Orav H. 1996. WORDNET: An On-line Lexical Database. *Papers of the First Swiss-Estonian Student Workshop on Computational and Theoretical Linguistics*: 64-68

WordNet 1.5 manuals (computer version)