[Translation provided by Krister Lindén of the Swedish orginal "FinnWordNet – WordNet på finska via översättning" appearing in LexicoNorcdica, vol 17, in 2010. Please cite the original.]

Krister Lindén, Lauri Carlson

FinnWordNet – Finnish WordNet by Translation

FinnWordNet is a WordNet for Finnish that conforms to the framework given in Fellbaum (1998) and Vossen (ed.) (1998). FinnWordNet is open source and currently contains 117,000 synsets. A classic WordNet consists of synsets, or sets of partial synonyms whose shared meaning is described and exemplified by a gloss, a common part of speech and a hyperonym. Synsets in a WordNet are arranged in hierarchical partial orderings according to semantic relations like hyponymy/hyperonymy. Together the gloss, part of speech and hyperonym fix the meaning of a word and constrain the possible translations of a word in a given synset. The Finnish group has opted for translating Princeton WordNet 3.0 synsets wholesale into Finnish by professional translators, because the translation process can be controlled with regard to quality, coverage, cost and speed of translation. The project was financed by FIN-CLARIN at the University of Helsinki. According to our preliminary evaluation, the translation process was diligent and the quality is on a par with the original Princeton WordNet.

1. Introduction

A WordNet is a thesaurus that is built on sets of synonyms with the same part of speech that can be interchanged in a given context. Such sets of synonyms are also called *synsets*. In WordNet the synsets have hierarchical partial orderings according to semantic relations, e.g. hyperonyms, hyponyms, meronyms, antonyms, etc. Each synset (with very few exceptions) has a gloss exemplifying or describing its meaning. A typical WordNet gloss includes a dictionary definition and an example, e.g. *devilize* 'turn into a devil or make devilish; "Man devilized by war". Some glosses give a hyperonym only (name of animal species for instance).

FinnWordNet is a WordNet for Finnish that conforms to the framework given in Fellbaum (1998) and Vossen (ed.) (1998) for the Princeton WordNet, which is open source and currently contains 117,000 synsets. FinnWordNet is a translation of the synsets in Princeton WordNet 3.0.

In Section 2, we give an overview of existing WordNets for other languages, how they have been created and some relevant figures on WordNet 3.0. In Section 3, we look at the benefits of manual translation, the challenges, the workflow and some of the practical work involved. In Section 4, we study some of the translation problems in theory and in practice and evaluate the translation work. In Section 5, we mention some of the work that remains now that the bulk translation is finished.

2 WordNets

WordNets in other languages have been created with various methods. We start with a brief overview of existing WordNets in other languages in Section 2.1, and then, in Section 2.2, we look at different approaches that have been used when creating or localizing a WordNet for another language. In Section 2.3, we assess the amount of work involved in creating a new WordNet from the Princeton WordNet.

2.1 WordNets for Other Languages

WordNets have been created for other languages than English, but generally they do not have as wide a coverage as the Princeton WordNet. In Table 1, we see a comparison of the sizes of different WordNets in different languages and language families. The figures have been retrieved from sources accessed on the internet in spring 2010.

WordNet	**Synsets**
Princeton WordNet (Fellbaum, 1998, WordNet 2010)	~120 000
EuroWordNet (Vossen, 2004a)	~10 000-50 000
BalkaNet (Vossen, 2004b)	~10 000
Polish WordNet (Piasecki & al., 2009)	~18 000
Danish WordNet (wordnet.dk, 2009; Bolette, 2010)	~41 000-60 000
Swedish WordNet (Viberg & al., 2002)	~15 000
NorNett (Fjeld & Nygaard, 2009) (80 000 rel.)	~50 000 synsets?

Table 1. WordNets in different languages or language families and their sizes in synsets

2.2 Creating a WordNet for a New Language

When creating a WordNet for a new language, there are three approaches to choose from: create a WordNet from scratch, translate a WordNet, or use a top ontology and extend it with a local synonym dictionary.

If, according to the first approach, a WordNet is created from scratch, the synsets have to be discovered as word senses from large corpora. A good example of this is Polish (Piasecki & al, 2009), where the synsets were discovered from large Polish corpora. However, in order to make sure that the most essential vocabulary was covered, they also used a core dictionary of Polish listing approximately 30 000 central words of Polish. In addition, a WordNet created from scratch

will, as a separate task, have to be linked to the Princeton WordNet in order to use it as a bilingual resource.

The second approach, i.e. translating the Princeton WordNet wholesale, forgoes any national pride claiming that a language is so different that it has to create its own synsets and synset hierarchies. From a language evolution perspective, the translation approach seems like a feasible idea, because most words are names of objects and phenomena in the world for which each language has invented or borrowed names. There may be a difference in granularity depending on the importance of a phenomenon in a language area, but the external world is largely the same for all cultures: the sun rises in the east, the water is wet and runs, we are born, become hungry, eat and die, etc. . It is only for more abstract concepts that one would expect significant divergence. One example is some specifically American concepts in the Princeton WordNet that do not exist with all their modern connotations in a potential target language. For some examples, see Table 2.

English	**Swedish**	**Finnish**
Hungry	hungrig	nälkäinen
Hunger	hungra	nähdä nälkää
Thanksgiving	skördefest?	kiitospäivä ≠ kekri

Table 2. Example of similarity and divergence of concrete vs. culture-specific concepts between languages. Hunger is global, but festivities vary.

A very common middle road and third approach to WordNet creation is therefore to translate, say, the top 5000 concepts of the Princeton WordNet in order to create a core WordNet and then extend this hierarchy with a local synonym dictionary. This is the approach taken by DanNet where they started with creating a local synonym dictionary based on a large Danish dictionary (Pedersen 2010). The intention is then to link DanNet to Princeton WordNet. This creates a link between the core concepts in the two languages and facilitates the use of the new WordNet as a bilingual resource. This approach was used in e.g. EuroWordNet (Vossen, 2004a). The Global Wordnet Association's selection of 5000 Base Concepts is based on structural properties of the wordnets in different languages, including concept position in the hierarchy and number or relations, concept frequency in the definitions or glosses, and morphological complexity and dependency of a concept. In many cases, sense frequencies are not available. As a result, the GWA list of 5000 base concepts is something

of a motley collection at the fringes (e.g. the American TV program *60 minutes* is included but *television* is not.).

On the one hand, creating a monolingual resource from scratch will end up largely replicating the decades of lexicographic work invested into the Princeton WordNet. On the other hand, translating the Princeton WordNet wholesale will have to deal with a number of American phenomena with no direct translation in the target language. Many languages already have synonym dictionaries and consequently many WordNet projects have used the third route, i.e. a combination of translation and a locally created sense hierarchy.

We believe that the basic unit of translation is a meaning unit, which can be represented as a synset with a hyperonym, a word class and a gloss as additional information to guide the translation of the synonyms in a synset. We therefore decided to translate all the synsets. This will give us parallel synsets for two languages so that we can reuse most of the semantic relations that have already been defined in Princeton WordNet and in addition get a bilingual dictionary.

Finnish as a language is unrelated to any of the languages with publicly available WordNets. In addition, we are unaware that a wholesale synset translation would have been attempted before. It remains to be evaluated in a separate phase how many missing purely Finnish words and word senses we may have to add.

2.3 Princeton WordNet 3.0

Below in Table 3 and 4, we give an overview of the size of the translation task we have chosen. In Table 3, we see that the Princeton WordNet (WordNet, 2010) has 117 659 synsets or meaning units. There are 155 287 distinct words (strings) in these synsets. Some words appear in several synsets. This creates a total of 206 941 senses, i.e. word-synset pairs, to translate.

Princeton WordNet	**Words**	**Synsets**	**Senses**
Noun	117 798	82 115	146 312
Verb	11 529	13 767	25 047
Adjective	21 479	18 156	30 002
Adverb	4 481	3 621	5 580
Totals	**155 287**	**117 659**	**206 941**

Table 3. Number of words, synsets and senses in the Princeton WordNet

In Table 4, we see that most of the words only have one meaning, i.e. appear only in one synset. As there are fewer synsets than words, it is to be expected that there is a sizable portion of synsets with more than one word in them. The average number of synonyms per synset is 1.8. On the other hand, some words (strings) appear with different senses in several different synsets, i.e. they are homonyms with different meanings. The average number of senses per word is 1.3. Most words only have one sense, so if we only consider the words with many senses, they will have nearly 3 senses on the average.

Princeton WordNet	**Monosemous Words**	**Polysemous Words**	**Homonymic Senses**
Noun	101 863	15 935	44 449
Verb	6 277	5 252	18 770
Adjective	16 503	4 976	14 399
Adverb	3 748	733	1 832
Totals	**128 391**	**26 896**	**79 450**

Table 4. Number of monosemous and polysemous words and homonymic senses in the Princeton WordNet. A word sense is homonymic if its string value appears in several synsets.

3. Manual Translation of Synsets

We opted for a professional manual translation of synsets of the Princeton WordNet into Finnish because the manual translation process is controlled in several ways:

1. Manual translation uses human intervention and therefore guarantees *high quality* compared with the machine-learning and self-organizing techniques that are normally used for word sense discovery from large corpora.
2. Wholesale translation guarantees *bulk*, i.e. approximately 200 000 word senses, so that we get more than yet another core vocabulary collection with only the core word senses.
3. Using a translation agency at a fixed cost per word guarantees *controlled cost***.**
4. A translation agency also guarantees *speed*, i.e. 3-6 months for the first version with up to 4-5 translators in parallel.

In Section 3.1, we look at some additional benefits of a translated WordNet. In Section 3.2, we look at the specific challenges of translating WordNet compared with translating other documents. In

Section 3.3, we outline the workflow and quality control of the translation process and, in Section 3.4, we look at some of the metadata that the translator is encouraged to provide in order to facilitate post-processing and further improvement of the translation.

3.1 Additional Benefits of a Bilingual WordNet

Translation gives us the ability to control the process and produce results efficiently, but in addition we will get a bilingual resource that is directly linked to the Princeton WordNet. Effectively, we will have a large freely available bilingual Finnish-English WordNet online within months and hopefully this can serve as an example for other languages as well.

The bilingual dictionary of parallel WordNets can be used for information retrieval (IR) and with some modifications also for machine translation (MT) due to its wide-coverage dictionary with approximately 100 000-200 000 sense disambiguated entries. It should be noted that key phrases in IR are not the most common words but the most salient words like names and terms, which are often missing from simple core bilingual dictionaries. It should also be noted that translations out of context are not directly useful for MT systems, but the English glosses and definitions can serve as a basis for further processing.

Our belief is that there are enough commonalities between the English and Finnish lexical resources to offset the differences. However, an evaluation will determine the extent of missing core Finnish concepts and the amount of semi-automated work needed to fill in missing concepts.

3.2 Translation Challenge Specific to WordNet

Can we make sure that a translator provides all the synonyms for a particular word sense and only synonyms with a given word sense? Our goal is partly in conflict with how professional translators are trained to translate a document, where it is preferable to translate with standard terminology and avoid non-standard variants. Now we are explicitly looking for variant translations in addition to the standard ones. The key to our solution was to give the translators the whole English synset to find Finnish equivalents for, i.e. the deal was to translate all the English synonyms in a synset subject to the meaning given in the gloss, using the closest equivalent for each synonym in the set. The gloss served as a guide for the translator to locate the correct equivalent in dictionaries and web sources. This induced the translators to find more than one

Finnish equivalent to match those given in English. In order to facilitate the translation process, we directed the translators to pay attention to:

1. The *gloss*, i.e. meaning context or word sense definition
2. The *hyperonym* of the word, i.e. conceptual abstraction level
3. The *part of speech*, i.e. morpho-syntactic context

We used a standard translation editor that supports XML formatting to protect the fields that we wish to show the translators for informational purposes in order to guide their translation of the synsets.

3.3 Workflow and Quality Control

Initially, we considered dividing the WordNet into sections according to subject area or topic, but not all terms have a subject field and many subject fields have relatively few terms, so we abandoned that idea in favor of a more pervasive and straightforward approach. We used the following workflow to enable quality control and an even distribution of labor among the translators:

1. The WordNet was transformed into XML format and divided into 20 alphabetical synset collections containing only the above-mentioned fields relevant to the translation work. About 10K orthographically identified proper name synsets (20K senses) were separated into a proper name file. Many less obvious named entities still remained in the other files.

2. A collection was handed off to a candidate translator, who first produced a translation sample for 1000 English synonyms.

3. The sample translation was inspected by a quality control translator.

4. If there were no major complaints, the translator was given further instructions on the basis of the sample and the whole collection was commissioned to be translated within a given deadline.

5. After submission, the synset collection went through a final quality control to determine whether the translation was within the expected quality bounds.

3.4 Practical Translation Work

To let the reader appreciate some of the translation process, we show a simplified partial sample of the XML encoding:

```
<GLOSS>an Egyptian descended from the ancient Egyptians
<HYPER ID="109700492">Egyptian
<SYNONYM PoS="n">
<Tuv Lang="EN-US">Copt
<Tuv Lang="FI">koptilainen
```

We chose to use SDL Trados as our translation environment because it is widely used in professional translation work. Trados TagEditor supports translation of XML documents. Trados keeps translated words in a translation memory. However, a translation memory offers little help in translating synsets, where occurrences of the same source word in different synsets are likely to need different translations. The editor helps restrict editing to the fields that are to be translated and allows us to define translation project specific additional tags. The translators use these tags to encode metadata about their translations, mainly to facilitate later maintenance and quality control.

After discussions with the translators, we converged on the following tags:

- `<or/>`
- `<approximate/>`
- `<broader/>`
- `<narrower/>`
- `<unconfirmed/>`
- `<note> … </note>`
- `<GEN/>`

The tag `<or/>` is used to separate alternative translations of a term to avoid object vs. meta-language ambiguity (punctuation like a comma can occur as part of an object language phrase). The tag `<approximate/>` could be used if the translator after a reasonable time could not locate an exact translation. The tag `<broader/>` indicates that the translation is broader than the original term and `<narrower/>` that the translation is narrower in scope than the original term. The purpose of these tags was not to problematize each and every match, but to allow translators to indicate lexical gaps. In practice, the usage rate of these tags was quite low (of the order of one in a thousand senses). Examples: hyperonym *nahka* ‘leather’ for *ooze leather,* hyponym *yläpurentainen* ‘overshot (jaw)’ for *overshot.* The tag `<unconfirmed/>` was to be used if the translator felt that a term was

accurate but could not confirm it in written sources. The tags <note> … </note> allowed the translators to add any information they saw fit to help later inspection and maintenance of the translations, for instance, to indicate difficulties in the English source.

The translators were given a translation manual ahead of the task with examples of good and avoidable practices. The quality control translator could then mostly just point to the manual to give further instructions. The translators were told to look for equivalents that could replace the source term in the context of the synset, avoid nonce translations, paraphrases, explanations of meaning, and avoid mixing object and metatext, such as parenthetical insertions, ad hoc grammar indications etc.

The translators were instructed to preserve part of speech when possible. Deviations from this were to be annotated to help subsequent automatic morphological analysis of the translations. The tag <GEN/> was provided for the common case of translating an English adjective satellite with a genitive noun in Finnish.

A fuller set of grammar annotations could have been developed, but was intentionally left out in order not to complicate the translators' task. Another traversal of the data by trained linguists can better attend to this aspect.

4. Translation Problems

Theoretically there may be several problems related to the translation of a thesaurus. In Section 4.1, we look at one instance of the prototypical problem of mismatching semantic fields. In Section 4.2, we turn to the problems observed in practice. In Section 4.3, we look at how to monitor the translation quality during the translation process and, in Section 4.4, we provide some general observations about the delivered translation quality.

4.1 Theoretical Challenge – Mismatching Semantic Fields

From a theoretical point of view, the major problem when translating between languages has been described as mismatching semantic fields of concepts. Concerning WordNet in particular, we may have mismatching synsets in two languages because not all of the words that are considered synonyms in one language are considered synonyms in another language. Our intuition is that this will mostly not be a problem for translation, because the gloss and the hyperonym restrict the domain

of translation, and the semantic space of the Princeton WordNet has been scrutinized and subdivided from many angles.

Let us have a look at one example where there is a mismatch between the semantic fields in English and Finnish due to the way Finnish and English treat their words for nationalities and languages. In both Finnish and English, the nationality is often a derived adjective of the name of a country. In Finnish, the name of the language often coincides with the name of the country, whereas in English, the language often coincides with the nationality.

```
<GLOSS> an Egyptian descended from the ancient Egyptians
<HYPER> Egyptian
<PoS="n" SENSE="1" COUNT="0">
      "EN-US": Copt
      "FI": koptilainen <or/> kopti

<GLOSS> the liturgical language of the Coptic Church used in
Egypt and Ethiopia; written in the Greek alphabet
<HYPER> Egyptian
<PoS="n" SENSE="1" COUNT="0">
      "EN-US": Coptic
      "FI": koptin kieli <or/> kopti

<GLOSS> of or relating to the Copts or their Church or
language or art; "the distinctive Coptic art of 6th-century
Christian Egypt"
<HYPER> Egyptian
<PoS="a" SENSE="1" COUNT="0">
      "EN-US": Coptic
      "FI": koptin kielinen <narrower/> <or/> koptilainen
```

In this case, the mesh of English WordNet is almost fine enough so that the translation of the concepts as they have been subdivided in WordNet causes no particular problems. This may of course be due to the fact that this is a whole group of words that display similar mismatches in many languages so they have long since been spotted and fixed.

We see as a major strength of the Princeton WordNet that it has undergone many years of maintenance and lexicographic work. That work gets taken advantage of in the translation approach. However on closer inspection, problematic areas remain.

A notable sore spot is the treatment of idioms in WordNet. Sometimes, they appear as separate meaning units with synonyms of their own, but not always. In a disturbing number of cases, WordNet splits a non-compositional part off the idiom and provides a separate synset for the part, like *tartaric* from *tartaric acid* or *take* from *take in*. This is a carryover from the traditional lemma based approach of paper

dictionaries. The best the translator can do is to translate the whole idiom and annotate the difficulty.

4.2 Problems in Practice

In practice, it turned out that it is sometimes difficult to find correct or even any translations for:

- medical terms, e.g. trade names of medicines
- new world plants and organisms
- chemical substances, e.g. trade names of active substances
- various isms, e.g. various religious movements
- legal terms
- business terms
- other terms specific to American culture and society

An example of untranslatable Americana is *hanging chad*, a perforation left hanging on a ballot ticket by a faulty voting machine – a front page news item in the 2000 elections. There are even 'WordNet words' which are difficult to find in the Websphere outside WordNet itself (or its many clones). Examples are the verb *spiritize* 'imbue with spirit', the nouns *spouter* 'a spouting whale', *tapper* 'a person who strikes a surface lightly and usually repeatedly', and the obscure terrorist group *Tareekh e Kasas* 'an organization of Muslims in India who killed Hindus in September 2002; believed to have ties with Muslim terrorists in Pakistan'. An improved online version of WordNet could perhaps keep some kind of usage statistics to age out ephemeral buzzwords and *hapax legomena.* On the other hand, these words are mostly a nuisance when translating WordNet which is done rather seldom. In daily WordNet usage, the odd words are not confused with the regular ones and cause no harm. In practice, it would probably cost more to remove them than the extra computer memory they require.

The odd words are examples of culture bound items or *realia* that have developed in American society during the last two hundred years and are local to American culture. Many are names for ideologies and societal phenomena that are unknown in Finnish culture and society. This seems to corroborate our initial hypothesis in Section 2.2 that only relatively new and culturally specific phenomena cause real translation problems. Fortunately, the problematic words are specific terms that rarely have synonyms. In addition, they are a small minority of the words in the WordNet and do not really motivate redoing the more than 95 % of unproblematic synsets from scratch.

A minor opposite defect is that WordNet sometimes makes sense distinctions that do not cause any translation difference in Finnish. Partly this could be due to an accidental similarity between English and Finnish.

4.3 Indicators of Translation Quality

In order to monitor the translation quality, we devised a set of indicators. We monitor each indicator by comparing a translator's output to the average for all translators. An alert in either direction brings the attention of the quality control translator to a particular synset. In addition to the statistics, the quality controller gets a sample of each type of outlier synsets sorted by decreasing degree of deviation.

The two main indicators are the number of variant translations per synset, and the numbers of untranslated or identically translated items. Other indicators follow the translators' use of annotation tags.

The indicator for variants monitors how many variants are provided in Finnish for each synset, compared with the number of variants in the English synset. Comparing this indicator with the cross-translator average, controls against undertranslation of synsets by repeating the same translation for all synonyms.

We allow that a word, after some reasonable consideration, may get no Finnish translation, which can be signaled by copying the English string as such. We want to monitor that this feature is not abused. We did this with an indicator for how many strings are identically translated to Finnish. In theory this should apply only to names with no need for transliteration, but for a small fixed price per word, a translator cannot be expected to do days of terminology research. If the rate of untranslated words is high for a translator, quality control needs to give the file some extra attention to determine the reason.

Similar indicators for how often tags were used on the average were also devised. An excessive number of variants is reflected in exceptional usage of the `<or/>` tag. Too many variants are sometimes an indication that the translator is unsure of the correct variant and is safeguarding himself by providing several options. The same is also true for excessively liberal or freewheeling translations reflected in an above-average number of times using the tags `<approximate/>` or `<unconfirmed/>`.

According to standard practice in the translation business, quality control should take less than 20 % of the translation time, so to speed up the quality control, process indicators were calculated as averages on all

synset collections to determine ratios for tag usage and synset sizes in English-to-Finnish translations. Collections with indicators deviating from the average were inspected more thoroughly. Synsets were rated against the averages of their synset collection. The top 500 deviating synsets of a collection were further inspected. While monitoring and inspecting the translations, we could find no deliberate attempts to misuse the process. Occasional mistakes that could be attributed to slips of attention were reported as general feedback to all translators.

4.4 Observations about the Delivered Translation Quality

Some general observations about the translation process can be made. First and foremost, we note that the translators were diligent and did their best as professional translators.

In the beginning, some translators tended to be overly cautious and more frequently used the tags `<unconfirmed/>` and `<approximate/>`, but their use was also cut down towards the end. Another strategy to cover occasional insecurity, in the beginning, was to provide an excessive number of synonyms. Outside proper names, less than 10 % of the strings were directly copied, Mostly copying was motivated by person names, trade names or other named entities without need for transliteration into Finnish.

In the beginning, translators tried to provide fairly complete sets of synonyms. Towards the end of their collections, they all tended to fall back into the normal single translation per sense translation mode. We can only speculate on the reasons for this, but one is surely the need to recover losses if too much time was spent on the initial part of the file.

This was not a surprise, as we were not counting on getting multiple translations per sense. The translators were paid to translate each sense at least once. We expected that to be enough to give us a significant rate of senses per word and synset.

Although we have only just begun to study the outcome more closely, we venture some preliminary figures here. ‘Finnish’ below includes only non-identical translations to Finnish (excluding proper names and other identically translated strings).

We interpret the figures in Table 5 as follows. About 180K Finnish strings translate about 210K English strings. The difference (about 30K) mainly consists of identically translated English names and other untranslatables. Only 287 senses were left untranslated. Most of them were numbers (the translation tool protected numbers from translation).

A little less than 100K different Finnish strings translate about 150K different English strings. The English figure 150K includes the identically translated strings. When the identical strings are subtracted from the comparison, we get 120K English strings, so the loss in lexical variety is small.

206723	English senses (tokens)
148556	English words (types)
1.39	English senses / word
180753	Finnish senses (tokens)
99639	Finnish words (types)
1.81	Finnish senses / word
.87	Finnish senses / English senses
.67	Finnish words / English words

Table 5. Comparison of numbers of synonyms for English and Finnish

Against the English Wordnet synset division, the Finnish translations are 1.3 times more ambiguous than the English originals, having 1.81 senses on the average against the original English 1.39. If you add the 30K identical strings here as well to the number of Finnish words, the number of senses per Finnish word is only 1.61 in Finnish compared with 1.39 in English. This indicates that slightly more Finnish words were reused in two or more synsets than was the case in English.

In order to give a taste of the lexical variety on the Finnish side, we compare the synonyms for *wave* in English with the synonyms of *aalto* 'wave' in Finnish. In WordNet, *wave* is a hyponym of *motion.* It gets two direct Finnish translations *aalto* and *laine*. Among its WordNet hyponyms are *billow, surge, comber, fluctuation, ripple, rippling, riffle, wavelet, roll, roller, rolling wave, seiche, surf, breaker, swell, crestless wave, whitecap, white horse.* These got about 10 Finnish translations (not listed here). A comparison with a Finnish synonym dictionary on the web comes up with nine more synonyms. Of these nine, six occur as translations of words under related but different hyperonyms *spray, breeze, wind, flow, foam*. Arguably, some of them do apply to waves but metonymously (e.g. *pärske* 'spume'). Of the proposals listed in the synonym dictionary, poetic *pärsky*, *lakkapää* and *vaahtopää* 'whitecap' had not occurred at all to the translators (*whitecap* was translated as *vaahtopäinen aalto* 'a white-capped wave'). On the other hand, the WordNet translation comes up with a spate of matter-of-fact hyponyms of *wave* not listed in the synonym dictionary, including *tsunami* and *backwash*.

5. Conclusion and Future Work

The timetable for the creation of the first version of the Finnish WordNet was strict. The translation process started in November 2009 and the raw translation was ready at the end of March 2010, in about 100 days[1]. If the translators had been able to maintain consistent translation speed of a thousand senses per day, this would have meant two full-time translators per day as a rough average. In practice, a translator's daily throughput varied between 500 and 1000 senses.

The need for further development will be determined by a search for missing Finnish base forms with an emphasis on high-frequency words. In addition, we need to locate missing senses of existing base forms, which probably needs to be done with semi-automatic methods in large corpora as in the work with the Polish WordNet (Piasecki et. al, 2009). A separate study will be carried out manually to determine whether there are parts of the translated WordNet with a missing or inadequate ontological structure for Finnish.

In order to correct mistakes in the translations and add new synset translations, the Finnish-English WordNet will be offered to the public as a large-scale free bilingual dictionary on the internet with the option for the general public to suggest corrections and improvements as a crowd-sourcing effort.

Acknowledgements

The project was financed by FIN-CLARIN at the University of Helsinki.

References

DanNet 2009: *DanNet – det danske wordnet*, http://wordnet.dk/ (March, 2010)

Fellbaum, Christiane 1998: *WordNet: An Electronic Lexical Database*. Ed. Cristiane Fellbaum, The MIT Press, Cambridge, London, England.

Fjeld, Ruth Vatvedt and Nygaard, Lars 2009: *NorNet — a monolingual wordnet of modern Norwegian*, Proceedings of the NODALIDA 2009 workshop WordNets and other Lexical Semantic Resources — between Lexical Semantics, Lexicography, Terminology and Formal Ontologies.

Piasecki, Maciej, Szpakowicz, Stanisław, and Broda, Bartosz 2009: *A Wordnet from the Ground Up*, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław.

Pedersen, Bolette 2010: Presentation at *Symposium om leksikografi og språkteknologi i Norden Schæffergården i Gentofte 29.-31.1.2010*

[1] Coinciding with the time frame for Finland's Winter War exactly 70 years earlier as noted by one of the translators, which most likely indicated her feelings about the excruciating translation effort.

Viberg, Åke, Lindmark, Kerstin, Lindvall, Ann, and Mellenius, Ingmarie 2002: *The Swedish WordNet Project***,** http://www.lingfil.uu.se/personal/viberg/SwedishWordNet2.pdf (March, 2010)
Vossen, Piek (ed.) 1998: *EuroWordNet: a multilingual database with lexical semantic networks for European Languages.* Kluwer, Dordrecht.
Vossen, Piek 2004a: *EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index.* Semi-special issue on multilingual databases. International Journal of Linguistics 17(2), June.
Vossen, Piek 2004b: Introduction, In *Romanian Journal of Information Science and Technology, Special Issue on Balkanet*, Volume 7, nr. 1-2.
Wordnet 2010: *WordNet – A lexical database for English*, http://wordnet.princeton.edu/ (March 2010)

Krister Lindén
Adjunct Professor, Ph.D.
University of Helsinki
Unionsgatan 40
FIN-00014 University of Helsinki
krister.linden@helsinki.fi

Lauri Carlson
Professor, Ph.D.
University of Helsinki
Unionsgatan 40
FIN-00014 University of Helsinki
lauri.carlson@helsinki.fi