Representing the Translation Relation in a Bilingual Wordnet

Jyrki Niemi, Krister Lindén

Language Technology, Department of Modern Languages, University of Helsinki
PO Box 24, FI-00014 University of Helsinki, Finland
{jyrki.niemi, krister.linden}@helsinki.fi

Abstract

This paper describes representing translations in the Finnish wordnet, FinnWordNet (FiWN), and constructing the FiWN database. FiWN was created by translating all the word senses of the Princeton WordNet (PWN) into Finnish and by joining the translations with the semantic and lexical relations of PWN extracted into a relational (database) format. The approach naturally resulted in a translation relation between PWN and FiWN. Unlike many other multilingual wordnets, the translation relation in FiWN is not primarily on the synset level, but on the level of an individual word sense, which allows more precise translation correspondences. This can easily be projected into a synset-level translation relation, used for linking with other wordnets, for example, via Core WordNet. Synset-level translations are also used as a default in the absence of word-sense translations. The FiWN data in the relational database can be converted to other formats. In the PWN database format, translations are attached to source-language words, allowing the implementation of a Web search interface also working as a bilingual dictionary. Another representation encodes the translation relation as a finite-state transducer.

Keywords: wordnet, bilingual representation, translation relation

1. Introduction

The Finnish wordnet, FinnWordNet (FiWN) is a translation of the word senses in the Princeton WordNet (PWN, version 3.0) (Fellbaum, 1998). This paper describes representations of the PWN–FiWN translation relation and the method of constructing the FiWN database by combining the structure (semantic and lexical relations) of the PWN and the Finnish translations of the word senses.

1.1. FinnWordNet as a Translation

Previous approaches to multilingual wordnets define the translation relation between senses (synonym sets, synsets), which we regard as too coarse for other than text-understanding purposes; for example, the word sense translation relation {*Amur* ↔ *Amur*, *Amur River* ↔ *Amurjoki*} would be reduced to the synset translation relation {*Amur*, *Amur River*} ↔ {*Amur*, *Amurjoki*}, losing the fine distinctions in the word sense translation correspondences. In this paper, we propose a solution with the translation relation defined between word senses.

The word senses of all the synsets of PWN were translated into Finnish by professional translators as described by Lindén and Carlson (2010). The direct translation approach is based on the assumption that most synsets in PWN represent language-independent real-world concepts. Thus also the semantic relations between synsets are assumed to be mostly language-independent, so the structure of PWN can be reused as well. This approach made it possible to create an extensive Finnish wordnet directly aligned with PWN, even though some words and concepts specific to the Finnish culture and society were left missing at this stage. FiWN currently contains 117,659 synsets, as does PWN. The translation relation also makes it possible to use FiWN linked with PWN as a bilingual English–Finnish–English dictionary, as well as a Finnish-only wordnet.

1.2. FinnWordNet Data Representations

Different uses of and different software for a wordnet require or benefit from different formats or representations of the wordnet data. Different formats include the PWN textual database format, relational database formats, such as the GermaNet database (Henrich and Hinrichs, 2010a), XML formats, such as Wordnet-LMF (Soria et al., 2009) and its extensions (Henrich and Hinrichs, 2010b), and simple plain-text lists of tab- or comma-separated values.[1]
A wordnet in the PWN database format can be searched using the PWN search software and various wordnet libraries; a relational database can be processed efficiently with standard database systems and libraries; a standardized XML format can be used for data interchange; and plain-text lists are often simple to use in rapidly-written scripts.
To ensure the consistency of the FiWN data in the different formats, we have chosen a relational (database) format as the primary format to which changes to the content are made and from which other formats are generated, either directly or via intermediate formats. Many common formats are monolingual, so we have extended them with a translation relation. We have conversion paths to the PWN database format, plain-text lists and finite-state transducers. Figure 1 shows the process of constructing FiWN, along with the different types of files and programs used. The FiWN data is freely available and can be downloaded in several formats from the FiWN project Web page.[2]
The Finnish translations of the PWN word senses as provided by the translators were in an XML format containing only a small part of the structure of PWN, serving as a translation context. To construct a usable wordnet for Finnish having the structure of PWN, we thus first combined the relations extracted from the PWN database with the Finnish translations of the PWN word senses.

[1] Tab- or comma-separated-values files can also be used as a plain-text representation of the tables of a relational database.

[2] http://www.ling.helsinki.fi/en/lt/research/finnwordnet/

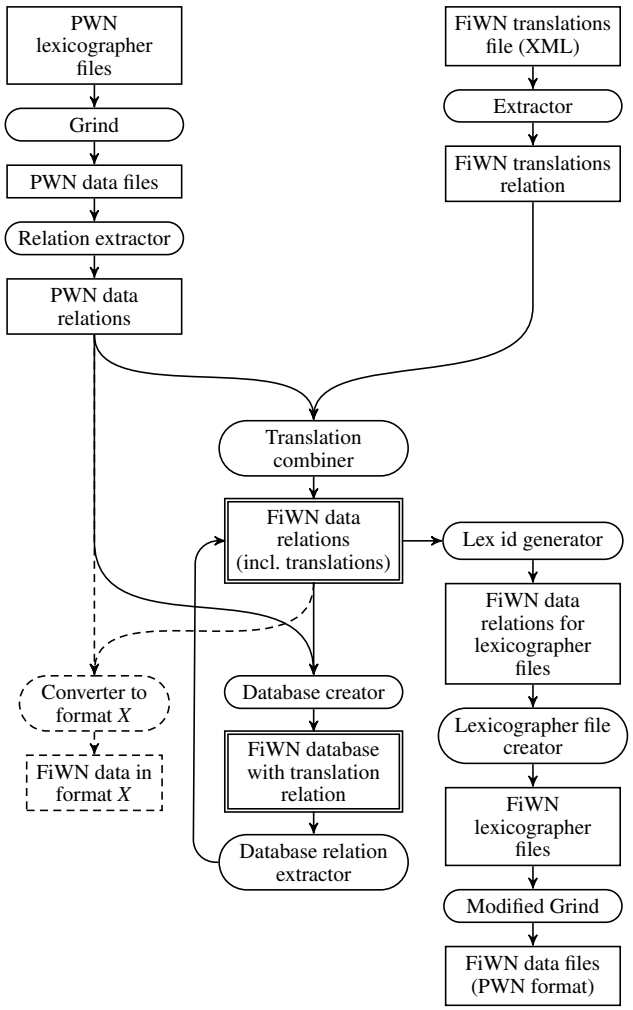

Figure 1: Constructing FiWN by combining the PWN structure and Finnish translations, with the result in multiple target formats. Rectangles denote files; double-bordered rectangles denote the main representations; rounded rectangles denote programs or scripts; dashed parts represent generic conversion paths to other target formats.

Since FiWN is a translation of the PWN word senses and retains the structure of PWN, we first converted the data to the PWN database format searchable with the PWN search software,[3] which is used in the FiWN Web search interface.[4] The major advantage of this approach was that we were able to develop the search interface more rapidly than otherwise. The PWN database format has no support for translations, but we attached them to word senses by using a special syntax processed by the search interface.

Since the PWN database format is very specialized and difficult to extend, we converted the FiWN data into a relational database format based on that used for the German wordnet, GermaNet (Henrich and Hinrichs, 2010a), modified and extended with word-sense and synset translation relations. Using the GermaNet database as the basis has the additional advantage that we could use an appropriately modified version of the GernEdiT editor (Henrich and Hinrichs, 2010a) to browse and edit FiWN data, which would make extending and modifying FiWN more user-friendly.

The rest of this paper is organized as follows. Section 2. describes the information content of the PWN database. Section 3. presents different ways of representing translations. Section 4. describes how we constructed the FiWN databases, and Section 5. how we succeeded in it. Section 6. discusses the process and directions for further work, and Section 8. concludes the paper.

2. Wordnet Information Content

The information in PWN is stored in the specialized PWN database format. To simplify constructing the FiWN database, we first extract the relations implicitly present in the PWN database.

2.1. Princeton WordNet Database

The PWN database uses a textual file format of its own, comprising data and index files, generated from the source *lexicographer files* by the PWN Grind program.[5]

Each part of speech has a separate data and index file. (Since we do not need the index files in constructing the FiWN database, we omit their description.) Each line in a data file contains the entry for a single synset with the information shown in Table 1. Figure 2 shows an example of a synset entry in a data file and the corresponding source entry in a lexicographer file. A detailed description of the database file formats is found in the PWN documentation.[6]

Offset of the line in the file
Source lexicographer file number
Synset type (generally, part of speech)
Number of words in the synset
List of words:
Word
Lexical identifier number (lex id)
Number of pointers (semantic and lexical relations)
List of pointers:
Relation type
Target synset offset
Part of speech
Source and target word number (for lexical relations)
Sentence frame numbers (for verbs only)
Gloss

Table 1: Information in a PWN data file entry (synset).

```
09833536 18 n 02 baker 0 bread_maker 0 002 @
10605985 n 0000 + 01663749 v 0101 | someone who
bakes bread or cake
```

```
{ [ baker, verb.creation:bake,+ ] bread_maker,
skilled_worker,@ (someone who bakes bread or
cake) }
```

Figure 2: PWN data file (data.noun) and lexicographer file (noun.person) entry for the synset {*baker, bread maker*}.

PWN synsets have no persistent identifiers. Within one PWN version, the synset offsets in a data file can be used as

[3] http://wordnet.princeton.edu/wordnet/man/wn.1WN.html
[4] http://www.ling.helsinki.fi/cgi-bin/fiwn/search
[5] http://wordnet.princeton.edu/wordnet/man/grind.1WN.html
[6] http://wordnet.princeton.edu/wordnet/documentation/

identifiers of a kind within a part of speech (e.g., 09833536 for {*baker, bread maker*} in Figure 2). The offsets change, however, if the length of the data in any synset changes. Alternative ways of referring to a synset typically refer to a representative of the synset, i.e., a word sense, by using, for example, a sense key of the PWN sense index file (e.g., *baker%1:18:00::* for {*baker, bread maker*}).[7]
The PWN database format is designed for efficient searching using binary search and random access. However, it would be difficult to use the database files *directly* as the basis for the FiWN database, as replacing English words with their Finnish translations would change file offsets and as a word may have several different translations. Instead, we first extract the information in the data files to intermediate relation files, which are then combined with the Finnish translations to FiWN relation files, from which other formats can be generated.

2.2. Extracted Relations

The relation files extracted from the PWN database represent the relations (in the sense of relational databases) implicitly present in the PWN database. The relation files are simple plain-text tab-separated-values files.
The information in the PWN database is divided into the following four relations (attributes comprising the primary key underlined):

SYNSETS(*synset_id*, *pos*, *lex_file*, *verb_frames*, *gloss*)
SEM_RELS(*source_synset_id*, *target_synset_id*, *rel_type*)

WORD_SENSES(*synset_id*, *lemma*, *lex_id*, *syn_marker*)
LEX_RELS(*source_synset_id*, *source_lemma*, *target_synset_id*, *target_lemma*, *rel_type*)

Verb frames only apply to verbs and the syntactic restriction marker (*syn_marker*) only to adjectives. The lexicographer file name (*lex_name*) of a synset and the lexical identifier (*lex_id*) of a word sense are only needed for generating lexicographer files; the lex id is used in lexicographer files to disambiguate between the different senses of a word.
Figure 3 shows the information extracted to the relation files from the the synset in Figure 2. Synset ids are in the format used in the FiWN translations file, described in Section 3.1., except that they are prefixed with a language code. The relation types use the PWN pointer symbols: @ denotes a hypernym and + a derivationally related form. Lexical relations contain the target lemma of the relation, which is not directly present in the synset entry but has to be retrieved from the target synset when extracting the relations.

3. Representing Translations

In this section, we present different representations of translations used for the FiWN data: the translations file produced by the translators, the translations extracted to a relation, the translation representations used in the PWN and relational database formats, synset translation linking via Core WordNet, and translation transducers.

[7] http://wordnet.princeton.edu/wordnet/man/senseidx.5WN.html

```
en:109833536|18||someone who bakes bread or cake
en:109833536|en:110605985|@
en:109833536|baker|0|
en:109833536|bread maker|0|
en:109833536|baker|en:201663749|bake|+
```

Figure 3: PWN data relations for the synset {*baker, bread maker*}: synsets, semantic relations, word senses and lexical relations (| denotes a tab separating fields).

3.1. FinnWordNet Translations File

The XML file containing the Finnish translations of the PWN word senses is organized by PWN synsets. The information for a synset contains a synset id, gloss and hypernym links, and for each synonym, its sense number, the English word in PWN and its Finnish translations. This information served as the translation context for the translators (Lindén and Carlson, 2010). Figure 4 shows a somewhat simplified sample entry in the translations file, corresponding to the synset in Figures 2 and 3.

```
<SYNSET ID="109833536">
  <GLOSS>someone who bakes bread or cake</GLOSS>
  <HYPERS>
    <HYPER ID="110605985">skilled_worker</HYPER>
  </HYPERS>
  <SYNONYM SENSE="2">
    <Tuv Lang="EN-US">baker</Tuv>
    <Tuv Lang="FI">leipuri</Tuv>
  </SYNONYM>
  <SYNONYM SENSE="1">
    <Tuv Lang="EN-US">bread maker</Tuv>
    <Tuv Lang="FI">leipuri<tai/>paakari</Tuv>
  </SYNONYM>
</SYNSET>
```

Figure 4: Translations file entry for the synset {*baker, bread maker*} (simplified).

The translations file uses as a synset id a nine-digit number consisting of the offset of the synset in the PWN 3.0 data files, prefixed with a part-of-speech code (1...4). In effect, the translations file identifies a word sense by the synset id and the English word; the sense number is redundant.
The translation relation is many-to-many: an English word sense may have several different Finnish translations, and several different English words in the same synset may have the same Finnish translation. Multiple translations of an English word sense are separated by the empty XML element *<tai/>* (meaning 'or'). Glosses were left untranslated because of resource constraints.
A number of the Finnish translations also contain translators' comments in the form of XML elements (Lindén and Carlson, 2010, 129–130). Comments are used to indicate that a translation is inexact, unconfirmed, broader or narrower than the original, or they may be free-form.

3.2. FinnWordNet Translations as a Relation

Some of the information in the translations file is unnecessary or redundant for constructing the FiWN database. To simplify further processing, we extract from the XML file a list of word senses containing the synset id, the English word and the corresponding Finnish translation. Moreover,

to make the file a more general translation mapping, we prefix the synset id with *en* for English and add another synset id for the translation, with prefix *fi* for Finnish.[8] We also convert the possible translators' comments to the type of the translation relation and extract free-form comments to a comment field. The relation type indicates whether a translation is exact or approximate, or broader or narrower than the original, corresponding to the EuroWordNet translation relation types *eq_synonym* (default), *eq_near_synonym*, *eq_has_hypernym* and *eq_has_hyponym* (Peters et al., 1998, 152–153). The result is a relation mapping English word senses to their Finnish translations:

TRANSLATIONS(*synset_id_en*, *lemma_en*, *synset_id_fi*, *lemma_fi*, *transl_type*, *comment*)

Figure 5 shows the translation relations extracted from the XML entry in Figure 4, with empty comments and = indicating *eq_synonym*.

```
en:109833536|baker|fi:109833536|leipuri|=|
en:109833536|bread maker|fi:109833536|paakari|=|
en:109833536|bread maker|fi:109833536|leipuri|=|
```

Figure 5: Translation relations for {*baker, bread maker*}.

3.3. Translations in the PWN Database

The PWN database and lexicographer file formats are rather rigid: they do not support adding other types of information than what is present in PWN (words, relation pointers, glosses and verb frames). They lack facilities for representing translations, comments or other information that we would like to attach to word senses or synsets.

Since some translations contained translators' comments as XML elements, we decided to represent the translations of a word sense in an XML element (*tr*) attached to the lemma. The Web search interface software recognizes these elements in the output of the PWN search tool and formats them appropriately, as shown in Figure 6.

```
Overview of noun leipuri

The noun leipuri has 2 senses (no senses from
tagged texts)

1. leipuri<tr>baker</tr> - (someone who bakes
commercially)
2. leipuri<tr>baker</tr>, leipuri<tr>bread
maker</tr>, paakari<tr>bread maker</tr> -
(someone who bakes bread or cake)
```

Overview of noun leipuri

The noun leipuri has 2 senses (no senses from tagged texts)

1. **leipuri** [baker] – (someone who bakes commercially)
2. **leipuri** [baker, bread maker], paakari [bread maker] – (someone who bakes bread or cake)

Figure 6: The output of the command `wn leipuri -over` using the Finnish–English database and the corresponding FiWN Web interface search result. An underlined word links to a search for the word.

The chosen representation of translations means that we have to generate separate versions of the FiWN database in the PWN data format for English–Finnish translations (PWN data with Finnish translations attached) and Finnish–English ones (Finnish translations with English original words attached).

3.4. Translations in the Relational Database

We use a modified version of the GermaNet database schema (Henrich and Hinrichs, 2010a, 2234–2235), extended with a translation relation. Although we regard the translation relation between word senses (*lexical units* in GermaNet) as primary, we also define a translation relation between synsets for compatibility with approaches aligning translations at the synset (semantic or *conceptual*) level.

We use separate tables for the translation relations instead of representing them as ordinary lexical or semantic relations, because they need two extra fields: *transl_type* and *comment*, used to encode the corresponding information in TRANSLATIONS; see Section 3.2.

The synset-level translation relation is inferred automatically from the lexical one by adding a translation relation between the synsets to which the words of a lexical translation relation belong; basically, it is the projection $\pi_{synset\_id\_en,\ synset\_id\_fi}(\text{TRANSLATIONS})$. In addition to that, we infer the synset-level translation relation type based on the translation relations of the word senses in the synset: if all the word senses have the same translation relation type, the synset translation relation will have the same type; otherwise, if any of them is an exact translation (*eq_synonym*), the synset translation is also; otherwise the synset translation is approximate (*eq_near_synonym*).

If a translation for a word sense is missing, we can use synset-level translations as a default: the translations of a word sense w_1 in synset S_1 are all the word senses w_2 in synset S_2 when S_2 is a translation of S_1.

Figure 7 shows a simplified outline of the database schema, with our extensions to the GermaNet schema in italics.

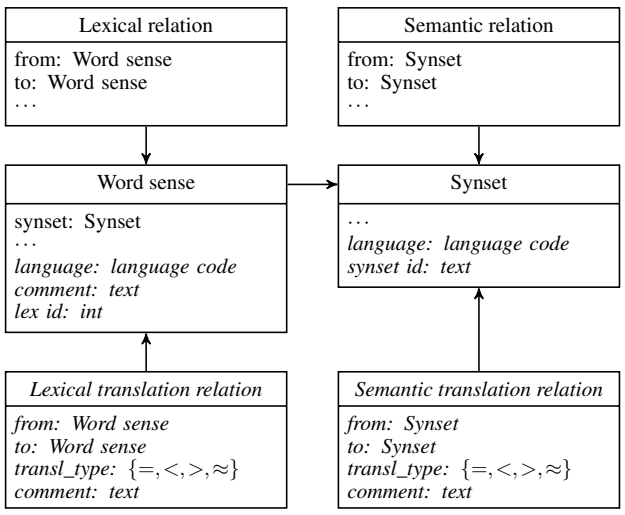

Figure 7: FiWN database schema (simplified), with the extensions to the GermaNet schema in italics.

We represent the synsets and word senses of both FiWN and PWN in the same tables. We thus need a language field in both the *Synset* and *Word sense* tables. Synsets

[8] For our purposes, it suffices to identify uniquely the synsets, so we also base the Finnish synset ids on the PWN synset offsets.

are language-specific to allow for different lexicalizations in different languages in the future.

The *synset id* field in the *Synset* table contains the FiWN or PWN synset id, which enables mapping the synsets back to PWN synsets. It is different from the internal id (key) in the *Synset* table. The *comment* field in *Word sense* may contain, for example, information on the register or usage of the word. The *lex id* field in *Word sense* is only used to save the lex id of the (PWN) word senses so that the data can be used to generate lexicographer files with the same lex ids as the original PWN data, thus also preserving the PWN sense keys for word senses.

3.5. Synset Translations via Core WordNet

One of the goals of the META-NORD project is to link the core senses of Danish, Estonian, Finnish, Norwegian and Swedish wordnets (Pedersen et al., 2012). As these wordnets vary in their content and format, they are linked via Core WordNet (CoreWN), consisting of almost 5,000 core senses of PWN. The approach resembles that of EuroWordNet, except that CoreWN is used instead of an interlingual index.

For each wordnet, a mapping relation is constructed for the senses (synsets) with correspondences to CoreWN senses. The mapping relation consists of CoreWN synset id (as a PWN sense key), the corresponding synset id of the other wordnet, translation relation type (the EuroWordNet types) and a weight. Since FiWN has direct word sense correspondences with PWN and thus also CoreWN, it is easy to construct the mapping between FiWN and CoreWN by projecting the translation relation onto synset ids and translation type, selecting the synsets present in CoreWN and adding a full weight.

The translation in language L_1 of a synset in the wordnet of language L_2 is generally retrieved via the intermediate CoreWN synset. Sometimes, however, the two languages may have only a non-exact translation relation to CoreWN, even though they would have an exact translation with each other. Such direct translation relations can be added separately to override the default indirect ones.

The resulting multilingual wordnet with translation relations defined between synsets could be stored using the database schema presented in Section 3.4., although only synset-level translations would be used.[9]

3.6. Translation Transducers

The PWN–FiWN translation relation can also be represented as a finite-state transducer (FST) that transduces an input word in one language to its translations in the other language, effectively functioning as a bilingual dictionary.

We have used the HFST tools (Lindén et al., 2009) to generate FSTs. The FiWN translation relation is first converted to the HFST-LexC format, which is then compiled into an FST. Composed with a morphological analyser of the source language, the FST recognizes inflected forms of words and translates them (with the base form of the target word). If synset ids are retained in the transduction, the result can be grouped by sense.

A demo of the translation FSTs is available.[10] The demo also includes synonym FSTs constructed from the FiWN data on the same principle; composed with a morphological generator, a synonym FST works as a thesaurus inflecting the output synonyms in the same form as the input word.

4. Constructing the FiWN Database

To construct the FiWN database, we join the data relations extracted from the PWN database files with the Finnish translation relation. From the resulting FiWN relations, we create a database in the PWN format (via lexicographer files) on the one hand, and a relational database on the other.

4.1. Combining Translations with the PWN Structure

The translation combiner program combines the FiWN translations relation with the PWN data relations files to create the corresponding FiWN data relations files. Since the synsets and semantic relations are transferred from PWN to FiWN as such, we only need to consider word senses and lexical relations.

Basically, the translation combiner joins the PWN word sense relation with the translation relation on the PWN synset id and the English word (lemma), and projects onto the synset id and the Finnish translations:

$$\text{WORD\_SENSES\_FI} = \pi_{\textit{synset\_id\_fi, lemma\_fi}}(\text{WORD\_SENSES\_EN} \bowtie_{\text{WS\_EN}.\textit{synset\_id}=\text{TR}.\textit{synset\_id\_en},\ \text{WS\_EN}.\textit{lemma}=\text{TR}.\textit{lemma\_en}} \text{TRANSLATIONS})$$

Lexical ids are generated by the lexicographer file creator.

The lexical relations relation is joined twice with the translation relation: once for the source synset id and lemma, and the second time for the target. From the result, we remove information unlikely to apply to Finnish: adjective syntactic restriction markers and verb frames. We also remove those derivationally related form and participle relations that are unlikely to be valid for the Finnish translations.

Figure 8 shows the FiWN data relations corresponding to the PWN data relations in Figure 3 combined with the translations in Figure 5.

```
fi:109833536|18||someone who bakes bread or cake
```

```
fi:109833536|fi:110605985|@
```

```
fi:109833536|leipuri||
fi:109833536|paakari||
```

```
fi:109833536|leipuri|fi:201663749|leipoa|+
```

Figure 8: FiWN data relations for the translation of the synset {*baker, bread maker*}: a synset, a semantic relation, word senses and a lexical relation.

4.2. Creating the FiWN Database in PWN Format

We generate the FiWN database in the PWN database format via lexicographer files, processed by a modified version of the PWN Grind program.

[9] The weight missing from the schema could easily be added.

[10] http://www.ling.helsinki.fi/cgi-bin/omor/omordemo.bash

To produce a FiWN database with translations attached to words, the lex id generator first optionally attaches translations from the FiWN translation relation to the FiWN or PWN word senses. It then generates lex ids for the resulting words. A lex id must be unique with regard to the combination of lexicographer file and the normalized form of the Finnish word that is used for indexing and searching the FiWN database in the PWN format.

The lexicographer file creator then creates lexicographer files based on the information in the FiWN data relations prepared for lexicographer files. The synsets are divided in lexicographer files in the same manner as in PWN.

The Grind program generates the PWN-format database from the lexicographer files. We have modified Grind to make it work with the FiWN data encoded in UTF-8. When normalizing words, the modified Grind removes all XML markup, in addition to lowercasing the words. The normalized words are used for searching, whereas the unnormalized words are stored in the data files as such. This enables us to search for a word and get its translation with it.

Figures 9 and 10 show examples of lexicographer and data file entries for the Finnish-only and Finnish–English FiWN database in the PWN format, respectively.

```
{ [leipuri, verb.creation:leipoa,+] paakari,
ammattitaitoinen_työntekijä,@ (someone who bakes
bread or cake) }
```

```
09647617 18 n 02 leipuri 0 paakari 0 002 @
10401982 n 0000 + 01575875 v 0101 | someone who
bakes bread or cake
```

Figure 9: Finnish-only lexicographer and data file for the translation of {*baker, bread maker*}.

```
{ [leipuri<tr>baker</tr>,
verb.creation:leipoa<tr>bake</tr>,+]
leipuri<tr>bread_maker</tr>2,
paakari<tr>bread_maker</tr>,
ammattitaitoinen_työntekijä<tr>skilled_
worker</tr>,@ (someone who bakes bread or cake) }
```

```
11803551 18 n 03 leipuri<tr>baker</tr>
0 leipuri<tr>bread_maker</tr> 2
paakari<tr>bread_maker</tr> 0 002 @ 12738550
n 0000 + 01907249 v 0101 | someone who bakes
bread or cake
```

Figure 10: Finnish–English lexicographer and data file entries for the translation of {*baker, bread maker*}.

4.3. Creating the FiWN Relational Database

Constructing the PWN/FiWN relational database is straightforward: the PWN data relations, FiWN data relations and the PWN–FiWN translation relation are converted to rows of the corresponding database tables and loaded into an appropriate database management system.[11] After that, appropriate indexes are created to improve the performance of the database.

[11] The first version of the database loader was written by Verena Henrich.

5. Testing the Data

Testing the FiWN data can be regarded as consisting of two parts: first, testing the technical correctness of the created database, and second, evaluating the quality of the translations themselves.

We have tested the correctness of the created FiWN database by randomly testing that the information originating from PWN (synsets, synonyms and relations) and the translations of the word senses have been correctly transferred to the database. For the PWN database format, we have used the command-line search tool and the Web interface based on it to test the database interactively.

Since the FiWN database is created in several steps, most of which convert data from one representation to another, each of the steps can be tested separately.

Based on an evaluation of the quality of the translations in FiWN, we estimated that only 0.5 ± 0.3 % of the evaluated translations should be replaced, which we deem to indicate a good quality of human translation (Niemi et al., 2012).

6. Discussion and Future Work

In general, we have found the different wordnet and translation representations useful for various tasks.

6.1. Translation Representations

The FiWN data model represents translations between word senses (or synsets) directly between two languages, without an interlingual index or other mediating structure. If the same database and translation relation is used to represent more than two languages, one language (typically English) can be chosen as an interlingua of a kind to overcome the explosion in the number of language pairs. A disadvantage of this approach is the partiality to the language chosen as the interlingua.

The different FiWN translation representations presented in Section 3. are suitable for different uses.

The translations file in XML format served well in the actual translation process by professional translators: the format was tailored to work with the Trados translation memory software used by the translators (Lindén and Carlson, 2010). It was not meant to be a complete representation of the structure of a wordnet but to serve as a translation context.

The relational representation of the translations was useful as an intermediate format for adding the translations to the FiWN databases. It was simpler to process than the XML and consistent with the other relation files. Joined with other information from the FiWN relations, this data has been used to verify selected translations manually and to correct them when appropriate.

Representing translations in the PWN database format enabled us to build the FiWN Web search interface on top of the PWN search software. The publicly available search interface has been used, for example, as an online thesaurus and a bilingual dictionary, to gather feedback on the synonyms and translations in FiWN, and to evaluate the usefulness of a wordnet for humans (Muhonen and Lindén, 2011). Instead of attaching translations in XML elements, any other well-defined representation could be used, as long as Grind is modified to recognize it.

The relational database format is a recent development, but we expect it to help us in expanding FiWN to make it a truly Finnish wordnet. It can also be used as the source format for other representations, either directly or via plain-text relation files.

Linking synset translations via Core WordNet is similarly new. It made it possible to link several wordnets built on different principles and using different formats, not all of them having word-sense correspondences with CoreWN.

The FST representation can be used as a stand-alone dictionary lookup that also recognizes inflected forms of the source-language word. Integrated with other software, such as a Web browser or a word processor, it might be helpful as a translator's tool. The dictionary FSTs are large, however, in particular if synset information is included.

6.2. Data Representations

The PWN database format is very specialized and difficult or impossible to extend without breaking compatibility with existing software. Moreover, data in the PWN database cannot be edited directly but only via the source lexicographer files, which are then compiled into the database.[12] Making even small changes to the data requires recompilation, and as noted by Henrich and Hinrichs (2010a), editing lexicographer files is laborious and error-prone. The lack of permanent identifiers for synsets also makes it more difficult to refer to a specific synset: using file offsets requires taking the data version into account, since when the data is changed, the synset offsets may also change.

In spite of all the above deficiencies, we find it justified to support the PWN database format as one of the data formats for FiWN, since it is supported by various pieces of software, even though they may lack the support for UTF-8-encoded data required for FiWN. Nevertheless, we think that the lexicographer files should be generated from a primary relational representation of the data.

A relational representation of data is conceptually simple and it can be operated on using well-defined operations. It is also extensible and portable, even though a physical representation of the relations in a relational database management system is not. For example, new types of semantic or lexical relations can easily be added.

In the relational representation, the pieces of information concerning a single entity, such as a synset, are scattered in several different relations. They could be gathered together in a structured representation of the data in an XML format. Other advantages of an XML format include standard (or industry-standard) means for processing data and possible human-readability. Because of the size of the FiWN data, however, processing the data in XML would very likely be slower than in a relational database. An XML format could nevertheless be useful as an interchange format.

Wordnet-LMF (Soria et al., 2009) is a "dialect" or a concrete XML instantiation for wordnets of the Lexical Markup Framework (LMF), a standard (ISO 24613) for lexical resource representation. For FiWN, Wordnet-LMF should be extended as proposed by Henrich and Hinrichs (2010b), to support lexical relations, although the other extensions would not currently be used for FiWN. Furthermore, Wordnet-LMF only has a translation relation between synsets, so it should be extended to cover also translations between word senses. The extension would appear easy using the same multilingual notations extension mechanism as for synset translations.

As long as we only add to FiWN new translations of PWN word senses, we could add them to the FiWN translations (XML or relation file) and combine them with the PWN data relations. However, when FiWN is extended beyond being a translation of PWN by adding synsets or semantic or lexical relations, the new data should be added to the FiWN data relations or the relational database, from which the data can be generated in other formats.

6.3. Improving FinnWordNet

Our aim is to extend and improve FiWN in several ways. We wish to add both new synonyms to existing synsets and completely new synsets to cover the current gaps. We are using Wikipedia and Wiktionary as sources of new synonyms for existing words (Niemi et al., 2012). We also plan to add missing frequent compound words as hyponyms of existing senses based on the final part of the compound (Pääkkö and Lindén, 2012). Although candidates for addition are produced semi-automatically and although they are also added to the relations in batch, we expect GernEdiT to be useful for manual verification.

7. Related Work

The EuroWordNet project (Vossen, 1998) combined several independent wordnets by linking them via an Inter-Lingual Index (ILI) (Peters et al., 1998; Vossen, 2004) to create a multilingual lexical database. The ILI is an unstructured list of concepts, to which the synsets in the individual wordnets are linked. A link from a synset to an ILI concept specifies that the synset represents a synonym, near-synonym, hypernym or hyponym of the ILI concept. The relations between the ILI concepts are provided by the wordnets. In addition, some language-independent relations are provided by a Top Concept ontology and domain labels, also linked to ILI concepts. As a further development, Fellbaum and Vossen (2008) outline Global WordNet linking different wordnets via a language-independent structured formal ontology.

The ILI approach explicitly allows for different lexicalizations of concepts in different languages, whereas FiWN at present mostly mirrors the English lexicalizations of PWN, even though for some word senses the FiWN translation relation contains the type of the translation derived from the translators' comments. Since EuroWordNet based ILI concepts on PWN (version 1.5) synsets, we expect that mapping FiWN synsets to ILI concepts would be fairly straightforward. The advantage of direct translation was that it allowed a relatively fast creation of a full-scale stand-alone wordnet for Finnish and its bilingual versions, which can be searched with the PWN search tool.

FiWN is more similar to the approach adopted in MultiWordNet (MWN) (Pianta et al., 2002), originally for the Italian wordnet, later also for several others. MWN links

[12] However, the free extJWNL Java library (http://extjwnl.sourceforge.net/) is promoted as being capable of directly modifying a wordnet database in the PWN format.

synsets directly to PWN synsets whenever possible and reuses the semantic relations of PWN, although they can be overridden if necessary. Lexical gaps can be explicitly represented by empty synsets and lexicalization differences by a *nearest* relation to a more general or more specific synset. In MWN, a translation relation is defined between synsets, whereas in FiWN it is between word senses.

8. Conclusion

FinnWordNet is a Finnish wordnet, currently a direct translation of the word senses of the Princeton WordNet. We believe that translations at the level of word senses are able to express finer distinctions of translation correspondences than those at the sense (synset) level.

The FiWN database was created by extracting the structure of PWN as relations and by joining with it the Finnish translations. The FiWN database has been generated both in the PWN database format and in a relational database format. We consider the successful creation of the FiWN databases as a proof of the technical correctness of the construction methods.

In the PWN database format, translations of word senses are attached to the original words, whereas in the relational database they are represented as a relation of their own. The database in the PWN format is used in the FiWN Web search interface. Future changes to the data will be made to the relational format, from which the PWN database can also be generated.

9. Acknowledgements

We thank our anonymous reviewers for their comments. We are grateful to Verena Henrich for fruitful cooperation in matters related to GermaNet and GernEdiT. This work was funded by the FIN-CLARIN (Finnish) and META-NORD (EC) projects.

10. References

Christiane Fellbaum and Piek Vossen. 2008. Challenges for a Global WordNet. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 75–82. City University of Hong Kong.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Global WordNet Association. 2012. *Proceedings of the 6th International Global Wordnet Conference (GWC 2012)*, Matsue, Japan.

Verena Henrich and Erhard Hinrichs. 2010a. GernEdiT – the GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2228–2235, Valletta, Malta. European Language Resources Association (ELRA).

Verena Henrich and Erhard Hinrichs. 2010b. Standardizing wordnets in the ISO standard LMF: Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 456–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17:119–140.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tools for morphology – an efficient open-source package for construction of morphological analyzers. In Cerstin Mahlow and Michael Piotrowski, editors, *State of the Art in Computational Morphology*, volume 41 of *Communications in Computer and Information Science*, pages 28–47. Springer, Berlin, Heidelberg.

Kristiina Muhonen and Krister Lindén. 2011. Do wordnets also improve human performance on NLP tasks? In Bolette Sandford Pedersen, Gunta Nešpore, and Inguna Skadiņa, editors, *Proceedings of the 18th Nordic Conference of Computational Linguistics NODALIDA 2011*, volume 11 of *NEALT Proceedings Series*, pages 146–152. Northern European Association for Language Technology (NEALT).

Jyrki Niemi, Krister Lindén, and Mirka Hyvärinen. 2012. Using a bilingual resource to add synonyms to a wordnet: FinnWordNet and Wikipedia as an example. In Global WordNet Association (2012), pages 227–231.

Paula Pääkkö and Krister Lindén. 2012. Finding a location for a new word in WordNet. In Global WordNet Association (2012), pages 286–293.

Bolette Sandford Pedersen, Lars Borin, Markus Forsberg, Krister Lindén, Heili Orav, and Eiríkur Rögnvaldsson. 2012. Linking and validating Nordic and Baltic wordnets: A multilingual action in META-NORD. In Global WordNet Association (2012), pages 254–260.

Wim Peters, Piek Vossen, Pedro Díez-Orzas, and Geert Adriaens. 1998. Cross-linguistic alignment of wordnets with an inter-lingual-index. In Vossen (1998), pages 149–179.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 international workshop on Intercultural collaboration*, pages 139–146, New York, NY, USA. ACM.

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Dordrecht.

Piek Vossen. 2004. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. *International Journal of Lexicography*, 17(2):161–173.