ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 163–172

Building a Wordnet for Turkish

Orhan BILGIN, Özlem ÇETINOĞLU, Kemal OFLAZER
Sabanci University, Istanbul, Turkey
E-mail: {orhanb,ozlemc,oflazer}@sabanciuniv.edu

Abstract. This paper summarizes the development process of a wordnet for Turkish as part of the Balkanet project. After discussing the basic methodological issues that had to be resolved during the course of the project, the paper presents the basic steps of the construction process in chronological order. Two applications using Turkish wordnet are summarized and links to resources for wordnet builders are provided at the end of the paper.

1. Introduction

The Human Language and Speech Technologies Laboratory at the Sabanci University, Istanbul, is a participant in the Balkanet Project [1]. The project basically follows the model of Princeton WordNet [2]. All six wordnets are linked to Princeton WordNet, to wordnets built under the EuroWordNet project [3] and to any other wordnet linked to the Interlingual Index (see EuroWordNet General Document [4], p. 39).

This paper summarizes the development process of Turkish wordnet. Section 2 presents a brief history of the construction process. Section 3 provides an overview of the current status of Turkish wordnet. Section 4 describes two applications which rely on Turkish wordnet. The paper concludes with a section containing practical tips, links to useful tools and resources, a basic reading list and a bibliography of papers on wordnets.

2. The Development Process

2.1. First Set of Concepts (Subset I)

At the beginning of the project, the Balkanet Consortium decided that each team translate the 1 310 Base Concepts of the EuroWordNet project (see EuroWordNet

General Document [4], p. 53). Base Concepts are important concepts that rank high in the hierarchy and have many hyponyms. Specifying an initial common set ensured maximum compatibility across the six Balkan wordnets.

2.2. Extracting Relations from Language Resources

After the translation of the first set of 1 310 synsets, the Turkish team made an attempt to automatically extract synonyms, antonyms and hyponyms from a machine-readable monolingual dictionary of Turkish:

Synonyms

The dictionary contained entries in the format $hw : w_1, w_2, ..., w_n$, where, hw is a headword and w_i is a single word. In these cases, the dictionary definition merely consists of a list of synonyms. This allowed us to extract almost 11,000 sets of potential synonyms, using a Perl script to parse dictionary entries.

There were also entry patterns like $hw : (w_i)*, w$. In these cases, a multi-word definition is provided after which a synonym is added, separated by a comma. These patterns gave us synsets in the form hw, w. A total of 10,846 such forms were extracted using a Perl script. These automatically extracted synonyms increased Turkish wordnet's average synset size from 1.20 to 1.35.

Hyperonmys

The existence of the phrases *"bir tür"* or *"bir çeşit"* (a kind of) in the dictionary definition indicates a hyperonymy relation between the headword and the lexical item that follows such phrases. 625 hyponym-hyperonym pairs were extracted in this way. In cases where the definition contained the phrase *"genel adı"* (general term for), more than one hyponym-hyperonym pair could be extracted from a single definition. Consider the following English example:

> Virtue: General term for moral qualities such as goodness, uprightness, integrity, morality.

81 such sets were extracted from the Turkish monolingual dictionary. The Turkish suffix *"-giller"* is usually used to construct taxonomic terms. Definitions of animals and plants usually contain this suffix, which allowed us to extract 889 hyponym-hyperonym pairs.

Antonyms

Existence of the word *"karşıtı"* (opposite of) in a dictionary definition indicates an antonymy relation between the headword and the lexical item preceding the word *"karşıtı"*. 235 antonym pairs were extracted in this way.

2.3. Second Set of Concepts (Subset II)

Having completed the translation of the first set of 1 310 synsets, the BalkaNet Consortium decided to expand the wordnets to 5 000 synsets during a second phase. Each team proposed a set of synsets, using various criteria (corpus frequencies, defining vocabularies, monolingual dictionaries, polysemy, etc.) to determine this new subset.

Selection of the Second Set in Turkish

While choosing the candidates for the second set, the Turkish team followed two different approaches. One of them was to find the so-called "gap synsets" and the other was to construct a set of candidates which would be used by all languages of the BalkaNet project. The resulting set of synsets has been formed by combining the results of these two approaches.

- **240 Gaps:** These are the 240 hyperonyms of Subset I which are not members of Subset I themselves. The idea here was to fill all the gaps between members of Subset I up to the relevant topmost entries in Princeton WordNet, so that the expanded set becomes a set of several "chains", and not "some chains and some free nodes" as it was before.
- **1 228 Additional Synsets:** While constructing this set of synsets, our aim was to obtain maximum "productivity" and maximum overlap between all languages. As a starting idea, we thought that the concept of a "defining vocabulary" was well suited to the task of determining "important" concepts. We used the defining vocabulary of the Longman Dictionary of Contemporary English which is freely available in the public domain. As a second source, we took the list of most frequent words in the English language, based on the British National Corpus. We found those entries in the Longman Defining Vocabulary which do not already exist among our extended set of synsets (1 310 from Subset I and the additional 240 "gaps") and we then found their intersection with the most frequent words of the English language. This intersection also allowed us to rank our new entries in terms of their frequencies, so, entries higher on the list could be considered more "important" than those lower. The result was a list of 712 lemmas. Then, we extracted all Princeton WordNet synsets that contain these lemmas. In this way, we obtained 3 114 synsets. Then, we reduced this set by taking only those synsets whose hyperonyms are Subset I synsets. The final product is a collection of 1 228 synsets. In this way, we eliminated all "dangling nodes" from our hierarchy. The resulting hierarchy contains 247 separate trees of varying length. This methodology is completely independent of the Turkish language. The motivation is that, at this relatively high level of the hierarchy, the most frequent words for any language would be "important" for all languages. In addition, the task we are faced with is the selection of synsets in the English language, since the translation was, for practical purposes, made from English. So, the idea was that basing the selection on English would not be a totally misleading method. Language-specific information gets more important as one moves down the hierarchy.

2.4. Shifting to Princeton WordNet 1.7.1

Before starting the translation of Subset II, the Consortium decided to shift from Princeton WordNet 1.5 to Princeton WordNet 1.7.1 as the basic resource. The aim was to avoid the problems involved in PWN 1.5.

2.5. Third Set of Concepts

After all partners finished the translation of Subset I and Subset II, the BalkaNet Consortium decided that all wordnets should reach 8 000 synsets at the end of a third phase. The Romanian team proposed that we take those synsets that exist in at least five EWN wordnets. The criteria of "avoiding gaps" was again applied.

2.6. Shifting to Princeton WordNet 2.0

During the translation of Subset III, Princeton University released WordNet 2.0 which contained thousands of additional synsets, verb groups, domain information for synsets and links between morphologically related items. Having observed that shifting from Version 1.7.1 to Version 2.0 would require minimal effort, the BalkaNet Consortium decided to shift to Princeton WordNet 2.0. Due to the structural changes introduced in Princeton WordNet 2.0, some synsets in Balkanet wordnets had to be merged, divided or deleted. The Czech team performed the transition automatically but some synsets had to be dealt with manually. Due to the shift to Princeton WordNet 2.0, the number of Base Concepts in the BalkaNet project is not equal to the number of Base Concepts in the EuroWordNet project.

2.7. Quality Validation

The following sections explain the quality validation tasks we have adopted:

Syntactic Quality

We first ensured the syntactic quality of Turkish wordnet in XML format. Each opening tag has a closing tag. All synsets have one and only one $< SYNSET >$ tag, one and only one $< ID >$ tag, one and only one $< POS >$ tag. Unless the synset corresponds to a lexical gap, it has at least one $< LITERAL >$ tag, together with its subtag $< SENSE >$. Lexical gaps have the special tag $< NL > yes < /NL >$.

We also checked the integrity of the values inside the tags. All values inside $< ID >$ tags are well-formed ILI numbers, i.e., are in the form ENG20-XXXXXXXX-Y, where X is a digit and Y is one of the letters n, v, a or b, representing parts of speech. This criterion also valid for the ILI numbers inside $< ILR >$ tags. All $< POS >$ tags have one of the following values: n, v, a or b. If a $< BCS >$ tag exists, it can only contain one of the following values: 1, 2, or 3, representing a synset in Subset I, Subset II and Subset III respectively. There are no empty tags such as $< LITERAL >< SENSE >< /SENSE >< /LITERAL >$ or $< DEF >< /DEF >$, etc.

Structural Quality

We obeyed the rule that the wordnets should not contain any "gaps". We avoided gaps by running a Perl script to find its gap hyperonmys in PWN 2.0 whenever we added a set of new synsets to our wordnet, and then, by translating all the gaps and adding them to the existing wordnet.

Closed-world Assumption and Dangling Relations

According to the closed-world assumption we adopted, if a relation is defined between ILI1 and ILI2, where ILI1 is a member of the wordnet, then ILI2 should also be a member of the wordnet. Relations where ILI2 is not contained in the wordnet are called "dangling relations". All such relations have been identified with the help of a Perl script and translated into Turkish.

Spelling Correction

The linguistic content of Turkish wordnet (synset members, glosses and usage examples, if any) has been passed through a spelling corrector. Although this may sound too simple and straightforward, it is a must if we consider that even Princeton WordNet 1.7.1 contains more than one thousand spelling mistakes.

Validation of Relations

All semantic relations imported from Princeton WordNet (or any other wordnet) have been manually, semi-automatically or automatically validated. For the synsets we translated, we imported all relations contained in PWN 2.0. In 95% of the cases, the semantic relations imported from Princeton WordNet made sense in Turkish.

Coverage Tests

Once the prototype wordnet is ready, it is a good idea to check its coverage of the lexical items in the local language. For a simple test, a simple wordlist can be used. For a more detailed analysis, one can use a wordlist with frequencies. Using a general-purpose corpus is another idea. Obviously, these methods only show you that a wordnet covers a certain percentage of the "word forms" and not "concepts" in your language. A practical idea could be to take a random sample and ask native speakers to decide whether or not a certain sense of a lexical item in the test corpus exists in the wordnet.

To measure the coverage of Turkish wordnet, we used two lists of unique words. The first wordlist was obtained from the Turkish translation of George Orwell's novel *1984* and the second wordlist was based on a general-purpose corpus. We ignored function words and passed the wordlists through a morphological analyzer and a POS tagger. The same procedure has been applied to all synset members. The intersection of Turkish wordnet with the *1984* corpus and the general-purpose corpus was found to be 87.40% and 86.45% respectively.

3. Current Status of Turkish Wordnet

Table 1 provides basic statistics on the current status of Turkish wordnet (March 2004). Table 2 shows the breakdown of Turkish wordnet's synsets into the three Balkanet Subsets and into parts of speech. Table 3 lists the semantic relations between synsets and gives the number of ocurrence of each relation.

Table 1. Basic Statistics

BASIC STATISTICS	NUMBER
Synsets	11 628
Synset Members	16 095
Average Synset Size	1.38
Unlexicalized Concepts	987
Definitions	4 913

Table 2. Distribution of Subsets and Parts of Speech

SYNSET TYPE	NUMBER	PERCENTAGE
Subset I	1 218	100%
Subset II	3 471	100%
Subset III	3 783	100%
Nouns	8 691	74.7%
Verbs	2 556	22.0%
Adjectives	381	3.3%

Table 3. Semantic Relations

RELATION TYPE	NUMBER
HYPERONYM	11 251
HOLO_MEMBER	946
HOLO_PART	1 423
HOLO_PORTION	176
CAUSES	100
BE_IN_STATE	577
NEAR_ANTONYM	1 400
SUBEVENT	127
ALSO_SEE	270
VERB_GROUP	896
CATEGORY_DOMAIN	384
TOTAL	17 550

4. Applications of Turkish Wordnet

The following two sections briefly describe two applications based on Turkish wordnet. The first application [5] is an attempt to export semantic relations to other wordnets based on purely morphological processes in the exporting language. The second application [6] is a wordnet-based “reverse dictionary” which provides words that correspond to a definition given by the user.

4.1. Using Turkish Wordnet to Export Semantic Relations to Other Wordnets

The basic idea is that morphological processes in a language can be effectively used to enrich individual wordnets with semantic relations. A more important claim is that morphological processes in a language can be used to discover less explicit semantic relations in other languages. This will both improve the internal connectivity of individual wordnets and also the overlap across different wordnets.

Using morphologically-related word pairs to discover semantic relations is by far faster and more reliable than building them from scratch. Morphology is a relatively regular and predictable surface phenomenon. It is a simple task to extract from a wordlist all instances which contain a certain affix, using regular expressions. Using morphological relations to discover semantic relations is a good way to start a wordnet from scratch or enrich an existing one.

A more interesting application of the method is the sharing of semantic information across wordnets. In the first case, which can be seen in Figure 1 semantically-related lexical items in both the exporting and the importing language are morphologically related to each other. Here, the importing language (Turkish in this case) could have discovered the semantic relation between *"deli"* (mad) and *"delilik"* (madness), for instance, by using its own morphology. So, the benefit of importing the relation from English is quite limited.

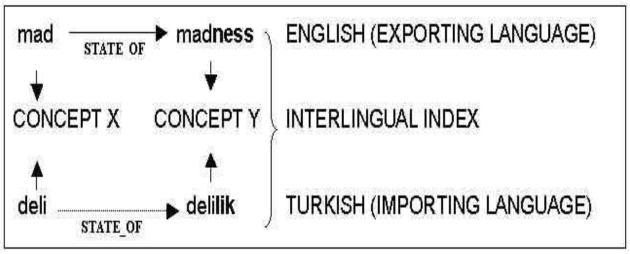

Fig. 1. Both languages involve morphology.

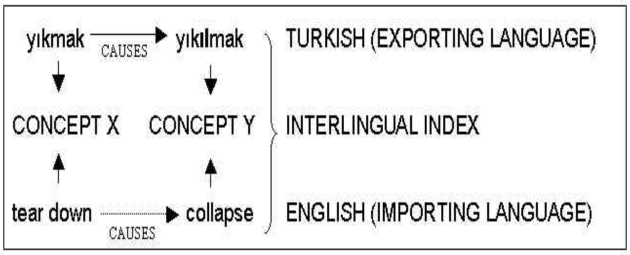

Fig. 2. Importing language does not involve morphology.

In the more interesting case, semantically-related lexical items in the importing language are not morphologically related to each other. For example, the causation relation between the lexical items *"yıkmak"* and *"yıkılmak"* is obvious to any native speaker (and morphological analyzer) of Turkish, while the corresponding causation relation between "tear down" and "collapse" is relatively more opaque and harder to discover for a native speaker of English and impossible for a morphological analyzer of English (Figure 2). Our method thus provides a way of enriching a wordnet with semantic information imported from another wordnet.

Some of the new links proposed involve morphologically unrelated lexical items which cannot be possibly linked to each other automatically or semi-automatically. Interesting examples in the case of the BECOME relation include pairs such as soap–saponify, good–improve, young–rejuvenate, weak–languish, lime–calcify, globular–conglobate, cheese–caseate, silent–hush, sparse–thin out, stone–petrify. Interesting examples in the case of the CAUSE relation include pairs such as dress–wear, dissuade–give up, abrade–wear away, encourage–take heart, vitrify–glaze.

4.2. Using Turkish Wordnet to Retrieve Words from Definitions

The application proposes a "Meaning-to-Word" system for Turkish, that finds a set of words closely matching the definition entered by the user. The approach of extracting words from "meanings" is based on checking the similarity between the user's definition and each entry of the Turkish database.

The system uses Turkish wordnet for query expansion. For example, suppose that the dictionary definition for *"bekar"* (bachelor) is *"evlenmemiş kimse"* (unmarried person). Although the user query *"evlenmemiş kişi"* (again, unmarried person) has an identical meaning, the meaning-to-word system fails to capture the almost 100% similarity between these two definitions. Using Turkish wordnet's synsets, the similarity between *"evlenmemiş kimse"* and *"evlenmemiş kişi"* becomes 100%.

The success rate of the proposed Meaning-to-Word system is measured in terms of the percentage of cases where the correct word is among the first 50 items suggested by the system. The use of the synonymy information in Turkish wordnet increased the system's success rate from 60% to 68%.

5. Tips and Resources

The following list provides practical tips and links to useful resources for wordnet builders. All of these methods and/or tools have in some way been used during the construction of Turkish wordnet:

- **VisDic** (`http://nlp.fi.muni.cz/projects/visdic`) is an emerging wordnet editor and viewer built by the Faculty of Computer Science at the Masaryk University in Brno, Czech Republic. Its latest versions are quite stable and can be used to build an entire wordnet from scratch. Visdic can be used to view any lexical database in XML format. This allows wordnet builders to access all lexical resources in a single application. Both Unix and Windows versions are available.
- **Google** (`http://www.google.com`) is one of the main tools of the lexicographer. Google provides purely empirical answers to lexical questions:
  - Up to five **asterisks** can be used in Google queries. Each asterisk corresponds to a word. For instance, one can find the correct equivalent of "verdict" in one's language by searching for "the court issued a *" and look for the words that correspond to the asterisk (obviously, the query has to be formed in the local language).

  - The **"define" function** of Google is a service which provides you with up-to-date definitions automatically harvested from the Internet by spiders (e.g. `define: "credit line"`).
  - **Google Images** (`http://www.google.com/imghp`) is a service which alleviates concerns regarding translational equivalence with its simple and ingenious methodology. To see what a "city bus" really is, one can just search it in Google Images and see if the pictures correspond to the translational equivalents in one's mind. Although one might think that this method can only be used for physical things, it sometimes produces good results even for lexical items as abstract as "ceremony".
  - Lexicographers use Google basically to compare the frequencies of two alternative words or phrases. `www.googlefight.com` facilitates this task by combining two Google searches into a single window.

- **Merriam-Webster Online** (`http://www.m-w.com`) is the ultimate authority on questions regarding the meanings of English words and phrases.

- **PERL** is probably the most suitable programming language for wordnet-related efforts. It can be freely downloaded at `http://www.activestate.com`.

- **Babylon** (`http://www.babylon.com`) is an excellent application containing hundreds of general-purpose and specialty dictionaries built by professionals and amateurs. The dictionaries can be used online or offline and the graphical system that allows you to lookup words without even typing them is quite successful. Babylon also has a downloadable copy of Princeton WordNet 2.0.

- `http://www.yourdictionary.com` is a **dictionary portal** where you can search for and access thousands of general-purpose and specialty dictionaries in hundreds of languages.

- Advanced text editors are part of the everyday life of wordnet builders. According to our experience, using **UltraEdit** (`http://www.ultraedit.com`) and **TextPad** (`http://www.textpad.com`) in combination produces the best results. TextPad's "Mark All" function is a perfect substitute for PERL scripts which extract matching lines from wordnet XML files.

- **Textpipe** (`http://www.crystalsoftware.com.au`) is a simple and powerful tool to extract information from huge text files. Lexicographers who do not know any programming language can use the simple graphical user interface to extract matches, count occurrences, join and split files, sort lines, etc.

- **Download Bot** (`http://www.sven-bader.de`) is a stable freeware for those who want to download huge amounts of data from the Internet. One can simply feed a list of 20 000 URLs and run the application overnight.

- **HTMASC** (`http://www.bitenbyte.com/htmasc.htm`) is a tool that removes tags from HTML files. After one downloads huge amounts of HTML-tagged text from the Internet using Download Bot, one can use HTMASC to convert them into clean text files and use Textpipe to massage them.

- **Teleport Pro** (`http://www.tenmax.com`) downloads entire websites. One can specify the file types one wants to download and the "depth" one wants the application to go. It is quite effective for building domain-specific corpora.
- It is a good idea to read Five Papers on WordNet [7] before starting to build a wordnet. The EuroWordNet project has produced several reports and deliverables. The EuroWordNet General Document [4] summarizes the results of the project in a single document.
  The Global Wordnet Association (`http://www.globalwordnet.org`) is a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world. Wordnet Bibliography, which is available at `http://engr.smu.edu/řada/wnb` is a good collection of papers on wordnets. The anthology pages of the Association for Computational Linguistics (`http://acl.ldc.upenn.edu`) contain a huge digital archive of research papers in computational linguistics.

References

[1] STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., *BalkaNet: A multilingual Semantic Network for Balkan Languages*, in *Proceedings of the First International WordNet Conference*, Mysore India, January 2002.

[2] FELLBAUM, C. (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[3] VOSSEN, P. (ed.), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht, 1998.

[4] VOSSEN, P., *EuroWordNet General Document*, `http://www.illc.uva.nl/EuroWordNet/docs.html` (Accessed: March 10, 2004).

[5] BILGIN, O., ÇETINOĞLU, Ö., OFLAZER, K., *Morphosemantic Relations In and Across Wordnets: A Study Based on Turkish*, in *Proceedings of the Global WordNet Conference*, Masaryk University, Brno, Czech Republic, January 2004.

[6] DURGAR-EL-KAHLOUT, I., OFLAZER, K., *Use of WordNet for Retrieving Words from their Meanings*, in *Proceedings of the Global WordNet Conference*, Masaryk University, Brno, Czech Republic, January 2004.

[7] BECKWIDTH, R., FELLBAUM, C., GROSS, D., MILLER, K., MILLER, G. A., TENGI, R., *Five Papers on WordNet*, Special Issue of the International Journal of Lexicography, **3**(4), 1990, 235–312.