# Chapter 15
# Turkish Wordnet

**Özlem Çetinoğlu, Orhan Bilgin, and Kemal Oflazer**

**Abstract** Turkish Wordnet is a lexical database for Turkish, built at Sabancı University in Istanbul, Turkey, between 2001 and 2004 as part of the Balkanet project. It currently contains 20,345 lexical items organized into 14,795 synonym sets (synsets hereafter), which are linked to each other via semantic relations such as hypernymy, antonymy, and meronymy. Turkish Wordnet uses the same concept pool as Princeton Wordnet, the eight wordnets of the Euro Wordnet project, and the five other wordnets of the Balkanet project. Synsets were added in several phases, starting with the most basic concepts at the top of the concept hierarchy. Monolingual resources were used to automatically extract semantic relations. Some semantic relations were extracted using the regular morphology of Turkish. Turkish Wordnet is available to researchers in the form of an XML file.

## 15.1 Introduction

This chapter provides an overview of Turkish Wordnet, a lexical database for Turkish, built at Sabancı University between 2001 and 2004 as part of the Balkanet project (Stamou et al. 2002), a 3-year, EU-funded project for the development of medium-sized wordnets for six languages: Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish.

Ö. Çetinoğlu
University of Stuttgart, Stuttgart, Germany
e-mail: ozlem@ims.uni-stuttgart.de

O. Bilgin
Zargan Ltd., Istanbul, Turkey
e-mail: orhan@zargan.com

K. Oflazer (✉)
Carnegie Mellon University Qatar, Doha-Education City, Qatar
e-mail: ko@cs.cmu.edu

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_15

A wordnet is an electronic lexical database where lexical items (words and phrases) are organized into synonym sets ("synsets"), each representing one underlying concept. Synsets are linked to other synsets by various semantic relations including hypernymy, meronymy, and antonymy. The original wordnet for the English language was built at Princeton University starting in 1990, and currently contains 155,287 unique lexical items grouped into 117,659 synsets (Fellbaum 1998). In response to the success of Princeton Wordnet, wordnets have been developed for more than 50 languages including Catalan, Chinese, Dutch, French, Greek, Hebrew, Hindi, Italian, Japanese, Kurdish, Persian, Russian, Spanish, and Turkish (Global Wordnet Association 2014).

During the 36-month Balkanet project, the Turkish team at the Human Language and Speech Technologies Laboratory of Sabancı University designed and developed a basic wordnet consisting of 20,345 lexical items organized into 14,795 synsets. The basic structure of Turkish Wordnet is largely based on Princeton Wordnet, and design decisions were jointly made by the Balkanet Consortium, of which the Turkish team was a member.

The following sections describe the design and development of Turkish Wordnet. We first provide an overview of the basic structure of Turkish Wordnet and then summarize the design decisions made by the Balkanet Consortium and the Turkish team. We provide basic statistics about the status of the wordnet as of the end of the project, and describe a series of validation tasks that were performed after the end of the development process to ensure consistency and quality. We then list work done by others that have utilized this resource and end with concluding remarks and some directions for future work.

## 15.2 Basic Structure of Turkish Wordnet

Like in all other wordnets built along the lines of Princeton Wordnet, the basic building block of Turkish Wordnet is a "synset," an abstract entity that acts as a container of lexical items (single words or multi-word phrases) which can be used to refer to the same concept in a given context. All lexical items that belong to the same synset have the same part of speech. Each synset has a unique identifier used to distinguish it from other synsets, a part-of-speech tag which is inherited by all synset members, and an optional definition (gloss) used to describe the concept the synset refers to.

### *15.2.1 Semantic Relations*

So far, the structure described above is not much different from a traditional thesaurus or synonym dictionary. What distinguishes a wordnet from these traditional language resources is that each synset can be linked to one or more other synsets to represent the semantic relations between the relevant concepts.

**Table 15.1** Semantic relations used in Turkish Wordnet

| Relation | Example |
|---|---|
| HYPERNYM | *kedi - hayvan* (cat - animal) |
| HOLO_MEMBER | *filo - deniz kuvvetleri* (fleet - navy) |
| HOLO_PART | *yarımküre - Dünya* (hemisphere - Earth) |
| HOLO_PORTION | *kar tanesi - kar* (snow flake - snow) |
| CAUSES | *koyulaştırmak - koyulaşmak* (to thicken (trans.) - to thicken (intrans.)) |
| BE_IN_STATE | *konforlu - konfor* (comfortable - comfort) |
| STATE_OF | *konfor - konforlu* (comfort - comfortable) |
| NEAR_ANTONYM | *iyi - kötü* (good - bad) |
| SUBEVENT | *horlamak - uyumak* (to snore - to sleep) |
| ALSO_SEE | *enerjik - aktif* (energetic - active) |
| VERB_GROUP | *hayal etmek - anlamak* (to imagine - to understand) |
| CATEGORY_DOMAIN | *mahkeme - hukuk* (court house - law) |
| SIMILAR_TO | *antidemokratik - otoriter* (undemocratic - authoritarian) |
| USAGE_DOMAIN | *Aspirin - marka* (Aspirin - brand) |

The HYPERNYM (or IS-A) relation is the basic semantic relation used to organize concepts into a hierarchical structure. For example, since a cat is a type of animal, the synset that the word *kedi* "cat" belongs to is linked to the synset that the word *hayvan* "animal" belongs to, via the HYPERNYM relation. Other major semantic relations used in Turkish Wordnet include NEAR_ANTONYM[1] (which links, for instance, *iyi* "good" to *kötü* "bad" to encode the antonymy relation), HOLO_PART (which links, for instance, *yumurta sarısı* "egg yolk" to *yumurta* "egg" to encode the part-whole relation), and CATEGORY_DOMAIN (which links, for instance, *mahkeme* "court house" to *hukuk* "law" to encode the fact that the concept of a court house belongs to the domain of law). Table 15.1 below lists all semantic relations used in Turkish Wordnet, along with examples.

### 15.2.2 Linking Wordnets to Each Other

Although an isolated wordnet in a single language can be a valuable resource in itself, it cannot be used in multilingual tasks such as cross-language search or

[1] Instead of the more straightforward relation name ANTONYM, Turkish Wordnet uses the name NEAR_ANTONYM to link two synsets with opposing meanings to each other. This is because antonymy is, strictly speaking, a relation that holds between individual *lexical items*, not between *concepts*. Consider the synset {*ascend, go up*}, where *ascend* and *go up* are synonyms in their relevant senses. But the two words have different antonyms: *descend* in the case of *ascend*, and *go down* in the case of *go up*. It would not be appropriate to link entire synsets to each other using the antonymy relation in its strict sense. Thus, along with the Euro Wordnet project (see Vossen (1998, p. 32)), Turkish Wordnet used the broader NEAR_ANTONYM relation to link synsets to each other.

machine translation unless two to more wordnets are mapped to each other. A simple way of mapping wordnets to each other is to use the same set of concepts, by using the same unique identifiers for the synsets.

This idea was first implemented by the Euro Wordnet project, which developed wordnets for Czech, Dutch, English, Estonian, French, German, Italian, and Spanish. All wordnets that were part of this project used the same set of synsets adopted from Princeton Wordnet 1.5. The so-called Inter-Lingual Index (ILI) assigns each synset in Princeton Wordnet a unique identifier based on the file offset of the relevant synset in the original Princeton Wordnet data files. This ensures that all eight wordnets of the Euro Wordnet project are connected to each other (see Vossen (1998, p. 39)).

The same method was adopted by the Balkanet project. Hence, Turkish Wordnet is perfectly mapped to Princeton Wordnet, the eight wordnets of the Euro Wordnet project, the five other wordnets of the Balkanet project, and any other wordnet that explicitly uses the concept pool of Princeton Wordnet. Figure 15.1 below depicts the basic structure of Turkish Wordnet as described above.

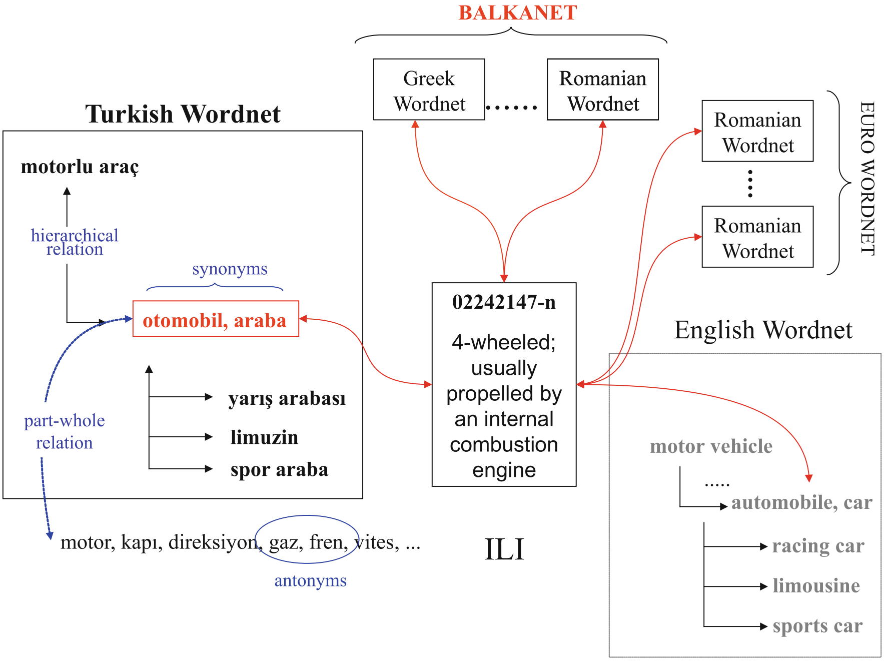

**Fig. 15.1** Basic structure of multiple wordnets. Turkish Wordnet is linked to other wordnets through the Inter-Lingual Index (ILI)

## 15.3 Design Decisions

In the development of the wordnets in the Balkanet Project, many decisions were made by the Balkanet Consortium and these were adopted by all six wordnet teams that were part of the project. Some were made locally by the Turkish team, taking into account the nature of the Turkish language and the tools and resources available. In both cases, however, the decisions were mainly based on the additional experience gained of the Princeton Wordnet and Euro Wordnet projects.

### *15.3.1 Merge vs. Expand*

Projects that aim to construct several interconnected wordnets usually prefer one of the two methodologies known as the "expand model" and the "merge model" in the literature (Vossen 1999). In the expand model, which is considerably simpler to implement, a fixed set of concepts (synsets) is taken from an existing wordnet, and each team translates the lexical items within these synsets into its local language. In the merge model, different synset collections built independently by each partner are combined into a single structure. The cost of the expand model is that the resulting wordnets are biased by the original wordnet, but the benefit is that all wordnets are linked to each other without extra effort. The benefit of the merge model is that the individual wordnets better reflect the structures of the individual languages, but the cost is the difficulty of combining the independent concept pools into a single, coherent structure. Since the Balkanet project aimed at maximum overlap with Princeton Wordnet and Euro Wordnets, it was decided at the outset to follow the expand model: Each team translated a fixed set of synsets from Princeton Wordnet 1.5.

### *15.3.2 Parts-of-Speech, Definitions, and Sense Numbers*

The Turkish Wordnet contains nouns, verbs, and adjectives. Considering that the project aimed at creating medium-sized wordnets covering the relatively more important synsets with the highest possible number of relations to other synsets, adverbs, which do not have a hierarchical structure and have relatively fewer semantic relations to other concepts, have not been included.

Another decision to be made is whether or not to provide brief definitions (glosses) for each synset. The Balkanet Consortium decided that this is a useful feature, and adopted it to the extent resources were available.

Since a given word or phrase can have several meanings, and can thus be a member of more than one synset, each word or phrase in a wordnet must have a unique sense number. The decision to be made at this point is whether sense numbers should be taken from an existing monolingual dictionary, or alternatively, be assigned automatically, and thus randomly. Since it is an extremely labor-intensive and error-prone task to map the senses of two separate lexicons to each other, we decided to assign sense numbers automatically. Consequently, unlike in a traditional monolingual dictionary where senses are ordered according to importance and frequency, sense numbers in Turkish Wordnet do not reflect an order.

### 15.3.3 Lexical Gaps

Each language organizes its lexicon in a different way than other languages. For example, English uses the single word "uncle" to denote one's father's or mother's brother, whereas Turkish uses the two different words *amca* and *dayı*, respectively, to cover the same conceptual space. This phenomenon, known as a "lexical gap," must be taken into account when designing a wordnet that shares its concept pool with one or more other wordnets. Since Turkish Wordnet is based on Princeton Wordnet, concepts that exist in Princeton Wordnet but are not lexicalized in Turkish create a problem. One option is to create "empty synsets" that have an ID, a position in the hierarchy, and even a part-of-speech tag, but no lexical content. Another possibility is to avoid empty synsets and provide EQ_HAS_HYPONYM and EQ_HAS_HYPERNYM links to the hyponyms and hypernyms of the lexical gap (see Euro Wordnet General Document (Vossen 1999, p. 38)). Turkish Wordnet has adopted the former approach. Currently, there are 1269 lexical gaps in Turkish Wordnet.

### 15.3.4 No Dangling Nodes or Relations

Another important design decision adopted in the Balkanet project is that, when a new synset taken from Princeton Wordnet is added to a local wordnet, all its hypernyms, up to and including a top node, have to be included in that local wordnet too. In other words, there should be no "dangling nodes"; it should be possible to reach a topmost node from any given node in the concept hierarchy.

As for semantic relations, since in the expand model adopted by Turkish Wordnet, all semantic relations were imported from Princeton Wordnet, adding a new synset to the wordnet resulted in the automatic addition of new semantic relations. In some cases, the synset(s) to which the new synset is linked through a semantic relation was already a part of the local wordnet. However, in some cases, there were "dangling relations" where the synset at one end of the semantic relation was missing. To avoid this, whenever an existing synset involved semantic relations that point to certain other synsets that were not yet part of the wordnet, such other synsets were also included in Turkish Wordnet.

### *15.3.5 Validating Semantic Relations*

When one imports semantic relations from another wordnet, it is possible that some of those relations do not hold between the lexical items of the importing language. Theoretically speaking, synsets correspond to concepts. So if everything goes well, there should be no cases where a semantic relation between those concepts ceases to be meaningful when expressed in another language. But there are certain reasons that give rise to such mismatches. First of all, “concepts” do not have an existence independent of a particular language. They have to be lexicalized. And every lexicalization inevitably involves some kind of attitude, culture, history, and pragmatic restriction. Secondly, concepts do not have well-defined boundaries. In many cases, a wordnet lexicographer translating a word or phrase from another language has to make do with a partial overlap of meaning. Therefore, each semantic relation imported from another wordnet has to be manually validated. The relations hold most of the time (around 95% according to our experience), but the few cases where they don’t hold were eliminated.

## 15.4 The Development Process

This section summarizes, in chronological order, the 36-month development process of Turkish Wordnet, based on the design decisions above adopted by the Balkanet Consortium and the Turkish team, both in view of the past experience of the Princeton Wordnet and Euro Wordnet projects, and the individual characteristics of the languages involved in the Balkanet project.

### *15.4.1 First Set of Concepts (Subset I)*

Having made the basic design decision of using the existing concept pool of Princeton Wordnet, the next logical step is to select an initial set of “important” concepts that will constitute the core of the new wordnet. One option would be to use local resources like a basic dictionary or word frequency list of Turkish. Although this would make sense from a monolingual point of view, the cost would be reduced overlap with other wordnets and increased difficulty of combining the six new wordnets that were being developed as part of the Balkanet project. In order to avoid these costs and maximize overlap with existing wordnets, the Balkanet Consortium decided that each team initially translate the 1310 “Base Concepts” (1010 nouns and 300 verbs) of the Euro Wordnet project (see Vossen (1999, p. 53)). Base concepts are concepts that rank high in the concept hierarchy and have the highest possible number of hyponyms.

### 15.4.2 Extracting Semantic Relations from Monolingual Resources

After the translation of the first set of 1310 synsets, we made an effort to increase average synset size by adding synonyms, and link these synsets to other synsets via the two basic semantic relations of hyponymy and antonymy. A machine-readable monolingual dictionary of Turkish (Türk Dil Kurumu 1983) was used to semiautomatically extract such relations.

#### 15.4.2.1 Synonyms

The monolingual Turkish dictionary we used for this purpose contained entries in the form $hw : w_1, w_2, \ldots, w_n$, where $hw$ is a headword and $w_i$ is a single word. In these cases, the dictionary definition merely consisted of a list of synonyms. This allowed us to extract 11,126 sets of potential synonyms, using a script to parse dictionary entries. The first row of Table 15.2 exemplifies a single-word definition that produces a two-member synset.

There were also entry patterns in the form $hw : w_1 w_2 \ldots w_n(, w)+$. In these cases, a multi-word definition is followed by one or more synonyms, 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 script. These automatically extracted synonyms were then filtered to cover the Base Concepts Subset I only and synonyms that were not already present in the existing synsets are selected. 196 such synset members were added to existing synsets, increasing Turkish Wordnet's average synset size from 1.20 to 1.35. The second row of Table 15.2 gives a definition where the last two words are the synonyms of the word *benzer* "similar" and produce a three-member synset.

**Table 15.2** Sample synsets automatically extracted from a Turkish monolingual dictionary

| Pattern | Example | Synset |
|---|---|---|
| $hw : w_1, w_2, \ldots, w_n$ | *Fonksiyon: işlev*<br>Function: role | {*işlev, fonksiyon*}<br>{role, function} |
| $hw : w_1 w_2 \ldots w_n(, w)+$ | *Benzer: Nitelik, görünüş ve yapı bakımından bir başkasına benzeyen veya ona eş olan, müşabih, mümasil*<br>Similar: That which resembles another in terms of appearance or structure, alike, homologous | {*benzer, müşabih, mümasil*}<br>{similar, alike, homologous} |

#### 15.4.2.2 Hypernyms

The existence of the phrases *bir tür* or *bir çeşit* "a kind of" in a dictionary definition potentially indicates a hypernymy relation between the headword and the lexical item that follows these phrases. 625 hyponym-hypernym pairs were extracted in this way. The first two rows of Table 15.3 show two such extractions.

In cases where the definition contains the phrase *genel adı* "general term for," more than one hyponym-hypernym pair can be extracted from a single definition. For example, four different hypernymy relations can be extracted from the definition in the third row of Table 15.3. A total of 81 such sets were extracted from the Turkish monolingual dictionary.

Finally, the Turkish suffix *-giller* "member of" is usually used to construct taxonomic terms. Definitions of animals and plants usually contain this suffix, which allowed us to extract 889 hyponym-hypernym pairs, as exemplified in the last row of Table 15.3.

**Table 15.3** Sample hypernym relations automatically extracted from a Turkish monolingual dictionary

| Pattern | Example | Hypernyms |
|---|---|---|
| *bir çeşit* | *barbut: Zarla oynanan **bir çeşit** kumar* | *barbut – kumar* |
| | Craps: **A kind of** gambling played with dices | Craps—gambling |
| *bir tür* | *vermut: Birçok bitkilerle özel koku verilmiş **bir tür** şarap* | *vermut – şarap* |
| | Vermouth: **A kind of** wine flavored with various herbs | Vermouth—wine |
| *genel adı* | *erdem: Ahlakın övdüğü iyilikçilik, alçakgönüllülük, yiğitlik,*<br>*doğruluk gibi niteliklerin **genel adı**, fazilet* | *iyilikçilik – erdem*<br>*alçakgönüllülük – erdem*<br>*yiğitlik – erdem*<br>*doğruluk – erdem* |
| | Virtue: **General term for** ethically praisable characteristics such as righteousness, integrity, purity, decency | Righteousness—virtue<br>Integrity—virtue<br>Purity—virtue<br>Decency—virtue |
| *-giller* | *mercimek: Bakla**giller**den, beyaz çiçekli bir tarım bitkisi (Lens culinaris)* | *mercimek – baklagil* |
| | Lentil: An agriculturally important **member of** legumes, having white flowers (Lens culinaris) | Lentil—legumes |

**Table 15.4** Sample near-antonym relations automatically extracted from a Turkish monolingual dictionary

| Pattern | Example | Near-antonyms |
|---|---|---|
| *karşıtı* | *çirkin: Göze veya kulağa hoş gelmeyen, güzel* ***karşıtı*** | *çirkin – güzel* |
| | Ugly: That which does not appeal to the eye or the ear, **opposite of** beautiful | Ugly—beautiful |
| *olmayan* | *temiz: Kirli, lekeli, pis, bulaşık olmayan* | *temiz – kirli*<br>*temiz – lekeli*<br>*temiz – pis*<br>*temiz – bulaşık* |
| | Clean: **That which is not** dirty, soiled, polluted, contaminated | Clean – dirty<br>Clean—soiled<br>Clean—polluted<br>Clean—contaminated |

#### 15.4.2.3 Near-Antonyms

Existence of the word *karşıtı* "opposite of" or *olmayan* "that which is not" in a dictionary definition indicates a potential antonymy relation between the headword and the lexical item preceding the words *karşıtı* or *olmayan*. In both cases, one or more near-antonyms can be derived from the definition. Table 15.4 shows a pair extracted from a definition including *karşıtı* and four pairs extracted from a definition including *olmayan*. A total of 235 antonym pairs were extracted in this way.

### *15.4.3 Second Set of Concepts (Subset II)*

Having completed the translation of the first set of 1310 synsets and having enriched these synsets using monolingual resources, the Balkanet Consortium then decided to expand the wordnets to 5000 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.

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

- **240 Missing Hypernyms:** These are the 240 hypernyms of Subset I synsets which are not members of Subset I themselves. The idea here is to fill all gaps between members of Subset I up to the relevant topmost nodes in Princeton

Wordnet, so that the expanded set becomes a set of several "chains," where it is always possible to reach a topmost node of Princeton Wordnet by moving up in the hierarchy.

- **1228 Additional Synsets:** While constructing this set of synsets, our aim was to choose concepts that are frequent, rank high on the concept hierarchy, are richly linked to other concepts, and would ensure maximum overlap between all languages represented in the project. As a starting idea, we proposed that the concept of a "defining vocabulary" was well suited to the task of determining such concepts. We used the defining vocabulary of the Longman Dictionary of Contemporary English (Quirk 1987). As a second source, we used the list of most frequent words in the English language, based on the British National Corpus (BNC Consortium 2001). We identified those entries in the Longman Defining Vocabulary which do not already exist among our extended set of synsets (1310 from Subset I and the additional 240 "missing hypernyms"), and we then found their intersection with the most frequent words of the English language. This intersection also allowed us to rank the 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 lexical items. We then extracted all Princeton Wordnet synsets that contain these lexical items, obtaining 3114 synsets. Then, we reduced this set by taking only those synsets whose hypernyms are Subset I synsets. The final product is a collection of 1228 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 of English 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 Balkanet wordnets were based on Princeton Wordnet. So, the idea was that basing the selection on English would not be misleading. The assumption is that language-specific information gets more important as one moves down the hierarchy.

### *15.4.4 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 certain problems involved in Princeton Wordnet 1.5, such as incorrect links, low-quality and missing glosses, and artificially divided synsets.

### 15.4.5 Third Set of Concepts (Subset III)

After all partners finished the translation of Subset I and Subset II, the Balkanet Consortium decided that all wordnets should reach 8000 synsets at the end of a third phase. It was decided that this phase should cover an additional 3000 synsets that exist in at least five Euro Wordnets. The criteria of "avoiding missing hypernyms" was again applied.

### 15.4.6 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 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, mostly automatically but sometimes also 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 Euro Wordnet project.

### 15.4.7 Adding Balkanet-Specific Concepts

Since the Balkanet project involved six languages from the Balkans and Eastern Europe, the expectation was that there existed a large number of regional/culture-specific concepts that the developers of Princeton Wordnet would not be expected to include in a wordnet of the English language.

Consequently, once the development of the core wordnets was finished using the existing concept pool of Princeton Wordnet, the consortium decided to shift to the "merge model." Initially, each team worked separately to develop its own set of language-specific concepts. The Turkish team developed 299 synsets, comprising 286 nouns, 10 verbs, and 3 adjectives. All Turkish synsets were equipped with brief definitions in English, and 141 synsets also had a picture. 285 of the Turkish synsets were linked to a Princeton Wordnet 2.0 synset via a hypernymy relation.

In the second step, all six teams came together to combine their individual contributions into a single repository called the "Balkanet Inter-Lingual Index" (BILI). The local synsets developed by each partner were checked by all the other partners; identical concepts were determined and assigned a single BILI number. The resulting set consisted of 332 Balkan specific synsets. As would be expected, most BILI concepts belong to culture- and region-specific domains such as the administrative system, religion, wedding traditions, architecture, food, animals, plants, traditional clothes, occupations, traditional arts, music, and tools. Some examples of the Turkish team's contribution are shown in Table 15.5.

**Table 15.5** Some language-specific concepts contributed by the Turkish team

| Lexical item | English definition |
|---|---|
| *incir reçeli* | Jam made of unripe wild figs. |
| *neyzen* | Person who plays the musical instrument *ney*. |
| *dayı* | Brother of one's mother. |
| *nazar boncuğu* | Charm made of blue, white, and yellow glass to protect you from the evil eye. |
| *mescit* | Small mosque where Friday prayers and special prayers on holy days are not held. |

### 15.4.8 Final Expansion

The purpose of this final expansion phase was to further increase coverage by adding those concepts that are frequently used in Turkish but do not yet exist in Turkish Wordnet. In order to determine these important and missing synsets, we took the 50,000 most frequent words of a 13-million-word in-house corpus compiled from six different domains of newspaper text. We then manually selected 2575 words that were decided to be important for Turkish and did not exist in Turkish Wordnet at that point.

This process was especially important for adjectives and certain closed classes such as cardinals, ordinals, and names of months, which were not represented in Turkish Wordnet.

## 15.5 Current Status of Turkish Wordnet

Table 15.6 provides basic statistics on Turkish Wordnet from the October 2014 release. The first three rows show the number of synsets, synset members, and average synset size. Note that average size is calculated simply as the ratio of synset members to synsets. It also includes those synsets that have zero members due to lexical gaps, which occur while trying to add an English synset to Turkish Wordnet via translation, as explained in Sect. 15.3.3 above. The current version of Turkish Wordnet contains 1269 such zero-member synsets. When these are ignored, average synset size rises to 1.50. 8792 of the synsets have only one member, while 3318 have two members, and 971 have three members. The two largest synsets of Turkish Wordnet have 10 members.

6717 of the synsets have a definition. 332 synsets have an SNOTE field that contains an English definition (for the Balkan-specific concepts). 141 of these English definitions are additionally associated with photos of concepts, with a SEE PICTURE identifier in the SNOTE field.

Table 15.7 shows the breakdown of Turkish Wordnet's synsets into the three Base Concept subsets, and into parts of speech. Note that the numbers of the original Base

**Table 15.6** Basic statistics on Turkish Wordnet

| Basic statistics | Number |
|---|---|
| Synsets | 14,795 |
| Synset members | 20,345 |
| Average synset size | 1.38 |
| Lexical gaps | 1269 |
| Definitions | 6717 |

**Table 15.7** Distribution of base concept subsets and parts of speech

| Synset type | Count | Part-of-speech | Count |
|---|---|---|---|
| Subset I | 1219 | Nouns | 11,227 |
| Subset II | 3470 | Verbs | 2736 |
| Subset III | 3782 | Adjectives | 792 |

**Table 15.8** Semantic relations

| Relation type | Number | Relation type | Number |
|---|---|---|---|
| HYPERNYM | 12,908 | CATEGORY_DOMAIN | 403 |
| SIMILAR_TO | 2497 | BE_IN_STATE | 327 |
| HOLO_PART | 1816 | STATE_OF | 290 |
| NEAR_ANTONYM | 1613 | HOLO_PORTION | 234 |
| HOLO_MEMBER | 1245 | CAUSES | 100 |
| ALSO_SEE | 1021 | SUBEVENT | 131 |
| VERB_GROUP | 923 | USAGE_DOMAIN | 32 |
| | | Total | 23,540 |

Concepts described in Sect. 15.4 do not match the numbers in the final version. This is due to the restructuring that occurred when we shifted from Princeton Wordnet 1.5 to 1.7.1, and then to 2.0. All three Base Concept subsets are 100% covered. As for the distribution of parts of speech, nouns dominate Turkish Wordnet with 75.9%, followed by verbs, which account for 18.6%, and adjectives, which constitute only 5.5%.

Table 15.8 lists the number of occurrences of each relation. Naturally, HYPERNYM is by far the most frequent relation. It is followed by SIMILAR_TO, HOLO_PART, and NEAR_ANTONYM. 7646 synsets have only one relation. 4077 of them have two, followed by 769 synsets with three relations. At the most highly connected end of the spectrum, there is one synset each with 29, 30, 32, 40, and 46 relations.

## 15.6 Quality Validation and Coverage Tests

Following the completion of the development phase, we performed a series of quality validation tasks. For the syntactic quality of the XML file, we used internally-developed scripts and the VisDic tool developed by the Czech team

(Horak and Smrz 2004b,a). VisDic, which is developed to visualize wordnets, also provides a set of tests for checking the consistency of wordnet XML files, such as duplicate IDs, duplicate lexical items, and duplicate links. The latter prevents a lexicographer from linking two synsets via more than one relation. For instance, a synset cannot be both the hypernym and antonym of another synset. A final VisDic test checks if the same lexical item with the same sense number occurs in more than one synset.

As for structural quality, we identified dangling nodes and dangling relations and added the respective missing synsets. We ensured all members of Base Concepts were present in the wordnet. In terms of content quality, we first passed the linguistic content of Turkish Wordnet (synset members, glosses, and usage examples, if any) through a spelling corrector. Then we manually, semiautomatically, or automatically validated all semantic relations imported from Princeton Wordnet. In 95% of the cases, the semantic relations imported from Princeton Wordnet were valid in Turkish as well.

As part of another major validation task, we measured the lexical coverage of Turkish Wordnet by checking the occurrence of high-frequency words of Turkish among synset members. The frequency word lists came from two different resources: The first one is a Turkish translation of George Orwell's novel *Nineteen Eighty-Four* and the second one is an in-house corpus.

While building the frequency lists, we morphologically analyzed and disambiguated all words using a morphological analyzer (Oflazer 1994) and applied the same procedure to synset members. While creating the list, we attached part-of-speech tags to the words, to avoid counting unmatching pairs as covered. As a result 76.6% of the *Nineteen Eighty-Four* lexical items were among synset members when we calculated the ratio of weighted sum of the successfully found lexical items to the total weighted sum. As expected, function words ranked high in the word list, and given that they were not included in the wordnet, they caused a reduction in the overall percentage. When we omitted function words, the percentage rose to 87.40%.

Similarly, we took the 50,000 most frequent words from the 13-million-word corpus mentioned above, excluding function words, and performed the same test. Coverage was 85.94%. When we considered the 20,000 most frequent words, it reached 86.45%. We then limited our list to the 1000 most frequent words of the corpus, and coverage rose to 87.32%.

## 15.7 Applications of Turkish Wordnet

This section provides an overview of projects and publications that are related to Turkish Wordnet and appeared either during or after the initial development phase.

### 15.7.1 Capturing Semantic Relations Through Morphology

The basic idea of this application is to effectively utilize morphological processes in a language to enrich individual wordnets with semantic relations. In a scenario where synsets of Wordnet A and Wordnet B are mapped to each other, simple morphological derivation processes in Language A can be used (1) to extract explicit semantic relations in Language A, and use these to enrich Wordnet A; (2) to verify existing semantic relations and detect mistakes in Wordnets A and B; and most importantly (3) to discover implicit semantic relations in Language B, and use these to enrich Wordnet B.

In this study, we focused on Turkish to extract morphological relations in the monolingual context, and propose relation extraction and verification both on Turkish and English in the multilingual context.

In the monolingual context, using morphologically-related word pairs to discover semantic relations is by far faster and more reliable than building them from scratch, especially in a morphologically-rich language with regular morphotactics. Productive affixes facilitate the derivation of lists of pairs using simple rules and improve the internal connectivity of a wordnet. In Bilgin et al. (2004), we identified 12 productive Turkish suffixes as candidates and proposed possible semantic relations for nine of them: WITH, WITHOUT, ACT_OF, ACQUIRE, MANNER, BECOME, BE_IN_STATE, CAUSES, PERTAINS_TO. Only the last three of these relations are defined in Princeton Wordnet and Euro Wordnet.

In the multilingual context, there are two cases: In the first case, semantically-related lexical items in both the exporting and the importing languages are morphologically related to each other, as can be seen in Fig. 15.2. Here, the importing language (Turkish) could have discovered the semantic relation between *deli* "mad" and *delilik* "madness," for instance, by using its own morphology. So, importing the relation from English does not bring an extra benefit. Yet, it can serve as a useful quality-control tool for the importing wordnet, and this has indeed been the case for Turkish.

Using the "expand model" in building Turkish Wordnet resulted in importing a set of relations together with the translated Princeton Wordnet synsets they belong to. Since Turkish employs a morphological process to encode, for example, BE_IN_STATE relations, the list of Turkish translation equivalents contains sev-

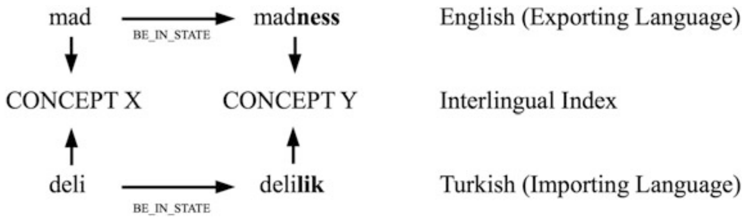

**Fig. 15.2** Both languages involve morphology

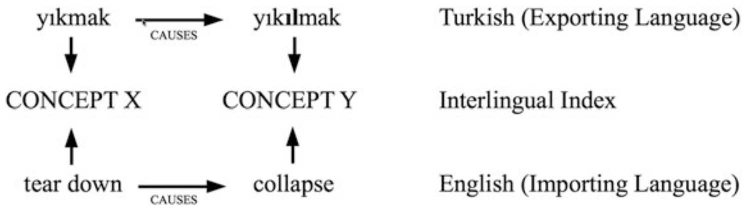

**Fig. 15.3** Importing language does not involve morphology

eral morphologically-related pairs like *deli-deli**lik*** “mad-mad**ness**,” *garip-garip**lik*** “weird-weird**ness**,” etc. Pairs that violate this pattern potentially involve incorrect translations or some other problem, and the translation method provides a way to detect such mistakes.

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 *yık**ıl**mak* 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 (Fig. 15.3). Our method thus provides a way of enriching a wordnet with semantic information imported from another wordnet.

We conducted a pilot study on two semantic relations to observe if this relation discovery procedure helps enrich Princeton Wordnet 2.0. We looked into the CAUSES (e.g., kill–die) and BECOME (e.g., stone–petrify) relations. CAUSES is a semantic relation that is present in Princeton Wordnet 2.0; BECOME on the other hand is not directly present, and is only represented by the underspecified relation ENG DERIVATIVE. 80 synset pairs in Turkish Wordnet have synset members related by a causative suffix that corresponds to the CAUSES relation. Only 18 of those pairs have a CAUSES relation in Princeton Wordnet 2.0. Similarly, 83 Turkish synsets were linked via the BECOME relation, by looking at the morphology of the lexical items. Only 11 of them were already linked in Princeton Wordnet 2.0.

Some of the new links proposed involve morphologically unrelated lexical items which cannot be possibly linked to each other automatically or semiautomatically. 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 CAUSES relation include pairs such as *dress-wear*, *dissuade-give up*, *abrade-wear away*, *encourage-take heart*, *vitrify-glaze*.

### 15.7.2 Turkish Wordnet in Use

Following its distribution, Turkish Wordnet has been used by several researchers as a basic lexico-semantic resource, either alone or in conjunction with another wordnet (usually Princeton Wordnet) or other NLP tools and resources.

Durgar-El Kahlout and Oflazer (2004) propose a meaning-to-word system for Turkish that finds a set of words matching the definition entered by the user. The system uses Turkish Wordnet to expand queries for the purpose of improving the coverage of matches. The use of the synonymy information in Turkish Wordnet increases the system's success rate from 60% to 68%. In another study, Durgar-El Kahlout and Oflazer (2005) take advantage of the links between Turkish Wordnet and Princeton Wordnet to construct a bilingual "root word alignment dictionary," which is used in the word-level alignment module of a statistical machine translation system. They report that the use of the wordnets significantly reduces noisy alignments. Oflazer et al (2006) report on the development of LingBrowser, a set of intelligent, active and interactive tools for helping linguistics students inquire and learn about lexical and syntactic properties of words and phrases in Turkish text. The system also incorporates information from Turkish Wordnet.

In his master's thesis, Boynueğri (2010) uses the definitions, synonyms, and semantic relations in Turkish Wordnet in a word-sense disambiguation task. Pembe and Say (2004) use synonymy information obtained from Turkish Wordnet to build the "lexico-semantic expansion" module of their linguistically motivated information retrieval system. Yücesoy and Öğüdücü (2007) use the hypernym hierarchy of Turkish Wordnet to propose an improved semantic similarity measure, which in turn is used in a document clustering task. Ambati et al. (2012) use the synsets of Turkish Wordnet to generate "coherent topics," which are then used to evaluate the performance of a "word sketches" system.

Özsert and Özgür (2013) use Turkish Wordnet in conjunction with Princeton Wordnet to improve the performance of their graph-based word polarity detection algorithm. The use of wordnets improves their accuracy from 84.5% to 95.5% in the case of Turkish, and from 91.1% to 92.8% in the case of English. In a related study, Demir (2014) uses semantic relations in Turkish Wordnet and Princeton Wordnet in a "valence shifting" task, which aims to "rewrite a text towards more/less positively or negatively slanted versions."

## 15.8 Conclusion and Directions for Future Work

In this chapter, we have described the design and development of Turkish Wordnet, a semantic network containing 20,345 lexical items organized into 14,795 synsets. Compared to Princeton Wordnet, Turkish Wordnet is a small-scale wordnet developed as part of an international project whose principal purpose was to produce six interconnected core wordnets that would also be linked to Princeton Wordnet and the eight wordnets of the Euro Wordnet project. Since an existing concept/relation pool in another language (English) was being used, the development process was largely a translation process, which had to be performed manually. During the 10 years that have elapsed since the conclusion of the Balkanet project, automatic and semiautomatic methods have been proposed and used for wordnet creation, which can inform future efforts to expand and enrich Turkish Wordnet.

Yıldırım and Yıldız (2012), for example, report the results of an experiment to automatically extract hypernym-hyponym pairs from a Turkish corpus, using lexico-syntactic patterns. Şerbetçi et al. (2011) extract a wider range of semantic relations from Turkish dictionary definitions, once again using lexico-syntactic patterns. They report having extracted more than 58,000 relations at 86.85% accuracy. These performance metrics suggest automatic or semiautomatic methods would facilitate inserting new synsets and relations to Turkish Wordnet.

The current version of Turkish Wordnet exclusively contains synset-to-synset relations, following decisions made on EuroWordNet. However, Princeton Wordnet has defined morphosemantic relations starting from version 2.0 (Miller and Fellbaum 2003) which establishes links between words that are connected to each other through derivational morphology. As explained in Sect. 15.7.1 above, the rich morphology of Turkish allows the automatic creation of a substantial number of word-to-word relations. Adopting Princeton Wordnet morphosemantic relations and using the proposed techniques in Bilgin et al. (2004) create an opportunity for the rapid automatic enrichment of Turkish Wordnet with semantic relations.

To summarize, we think that future efforts to expand and enrich Turkish Wordnet might benefit from automatic and semiautomatic methods that rely more on language resources in Turkish and specific features and mechanisms that are peculiar to the Turkish language.

The XML distribution of Turkish Wordnet is available for research purposes at bitbucket.org/ozlemc/twn/ (Accessed Sept. 14, 2017), together with the VisDic configuration files to visualize and edit the wordnet.

## References

Ambati BR, Reddy S, Kilgarriff A (2012) Word sketches for Turkish. In: Proceedings of LREC, Istanbul, pp 2945–2950

Bilgin O, Çetinoğlu Ö, Oflazer K (2004) Morphosemantic relations in and across wordnets: a preliminary study based on Turkish. In: Proceedings of the second global WordNet conference, Brno, pp 60–66

BNC Consortium (2001) British national corpus. www.natcorp.ox.ac.uk/. Accessed 3 July 2017

Boynueğri A (2010) Cross-lingual information retrieval on Turkish and English texts. Master's thesis, Middle East Technical University, Ankara

Demir Ş (2014) Generating valence shifted Turkish sentences. In: Proceedings of the eighth international natural language generation conference, Philadelphia, PA, pp 128–132

Durgar-El Kahlout İ, Oflazer K (2004) Use of wordnet for retrieving words from their meanings. In: Proceedings of the second global WordNet conference, Brno, pp 118–123

Durgar-El Kahlout İ, Oflazer K (2005) Aligning Turkish and English parallel texts for statistical machine translation. In: Proceedings of ISCIS, Istanbul, pp 616–625

Fellbaum C (1998) WordNet: an electronic lexical database. MIT Press, Cambridge, MA

Global Wordnet Association (2014) Wordnets in the world. www.globalwordnet.org/wordnets-in-the-world. Accessed 3 July 2017

Horak A, Smrz P (2004a) New features of wordnet editor VisDic. Rom J Inf Sci Technol 7(1–2):1–13

Horak A, Smrz P (2004b) VisDic- wordnet browsing and editing tool. In: Proceedings of the global WordNet conference, Brno, pp 136–141

Miller GA, Fellbaum C (2003) Morphosemantic links in wordnet. Traitement Automatique de Langue 44(2):69–80

Oflazer K (1994) Two-level description of Turkish morphology. Lit Linguist Comput 9(2):137–148

Oflazer K, Erbaş MD, Erdoğmuş M (2006) Using finite state technology in a tool for linguistic exploration. In: Proceedings of FSMNLP, Helsinki, pp 191–202

Özsert CM, Özgür A (2013) Word polarity detection using a multilingual approach. In: Proceedings of CICLING, Samos, pp 75–82

Pembe FC, Say ACC (2004) A linguistically motivated information retrieval system for Turkish. In: Proceedings of ISCIS, Kemer, pp 741–750

Quirk R (1987) The Longman American defining vocabulary. www.longmandictionariesusa.com/res/shared/vocab_definitions.pdf. Accessed 3 July 2017

Şerbetçi A, Orhan Z, Pehlivan İ (2011) Extraction of semantic word relations in Turkish from dictionary definitions. In: Proceedings of the workshop on relational models of semantics, Portland, OR, pp 11–18

Stamou S, Oflazer K, Pala K, Christoudoulakis D, Cristea D, Tufis D, Koeva S, Totkov G, Dutoit D, Grigoriadou M (2002) Balkanet: a multilingual semantic network for Balkan languages. In: Proceedings of the first global WordNet conference, Mysore

Türk Dil Kurumu (1983) Türkçe Sözlük. Türk Dil Kurumu, Ankara

Vossen P (ed) (1998) Euro WordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht

Vossen P (1999) EuroWordNet general document. www.vossen.info/docs/2002/EWNGeneral.pdf. Accessed 3 July 2017

Yıldırım S, Yıldız T (2012) Automatic extraction of Turkish hypernym-hyponym pairs from large corpus. In: Proceedings of COLING, Mumbai, pp 493–500

Yücesoy B, Öğüdücü ŞG (2007) Comparison of semantic and single term similarity measures for clustering Turkish documents. In: Proceedings of the international conference on machine learning and applications, Cincinnati, OH