# Morphosemantic Relations In and Across Wordnets
## A Study Based on Turkish

Orhan Bilgin, Özlem Çetinoğlu, and Kemal Oflazer

Sabanci University, Human Language and Speech Technologies Laboratory
Istanbul, Turkey
Email: orhanb@sabanciuniv.edu, ozlemc@sabanciuniv.edu,
oflazer@sabanciuniv.edu

**Abstract.** Morphological processes in a language can be effectively used to enrich individual wordnets with semantic relations. More importantly, 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 morphology to improve the quality of wordnets and to automatically prepare synset glosses are two other possible applications.

## 1 Introduction

Over the recent years, wordnets have become important resources for natural language processing. The success of Princeton WordNet (PWN) [1] has motivated the development of several other wordnets for numerous other languages[1].

Wordnets are lexical semantic networks built around the concept of a 'synset', a set of lexical items which are synonymous in a certain context. Semantic relations such as hyperonymy, meronymy and antonymy link synsets to each other and it is these semantic relations that give wordnets their essential value.

The number of semantic relations among synsets is an important criterion of a wordnet's quality and functionality. Thus, any method that would facilitate the encoding of semantic relations will be greatly helpful for wordnet builders. Furthermore, the recent proliferation of wordnets opened up the possibility of cross-linking across wordnets.

In this paper, we claim, with special emphasis on Turkish, that morphological processes in individual languages offer a good starting point for building wordnets and enriching them with semantic information encoded in other wordnets[2].

The present paper is structured as follows: Section 2 describes possible applications of the proposed method in monolingual and multilingual contexts. Section 3 provides an overview of the methodology in language-independent terms. Section 4 clarifies this methodology further by providing a case study for Turkish morphology and the possibility of exporting

[1] See "Wordnets in the World" at http://www.globalwordnet.org/

[2] The exchange of semantic relations across languages requires that the importing wordnet and the exporting wordnet are linked to each other in some way. The EuroWordNet project [2] and the BalkaNet project [3] solved this by introducing the concept of an 'Interlingual Index' (ILI), a common repository of language-independent concepts to which all other languages would be linked.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 60–66.

semantic relations from Turkish into English. Section 5 draws conclusions and provides insights regarding possible future work.

## 2 Areas of Application

Possible applications of the methodology proposed in this paper can be more formally described as follows:

Simple morphological derivation processes in a certain Language A can be used (i) to extract explicit semantic relations in Language A and use these to enrich Wordnet A; (ii) to detect mistakes in Wordnet B; (iii) to automatically prepare machine-tractable synset glosses in Wordnet A and/or Wordnet B; and most importantly (iv) to discover implicit semantic relations in Language B and use these to enrich Wordnet B.

The following three subsections clarify these applications in monolingual and multilingual contexts.

### 2.1 Monolingual Context: Single, Isolated Wordnet

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.

### 2.2 Multilingual Context: Several Wordnets Linked to Each Other

The more interesting application of the method is the sharing of semantic information across wordnets. There are two cases: i. Semantically-related lexical items in both the exporting and the importing language are morphologically related to each other.

In this case, the importing language (Turkish) could have discovered the semantic relation between “deli” (mad) and “deli**lik**” (mad**ness**), for instance, by using its own morphology. So, the benefit of importing the relation from English is quite limited. Still, importation can serve as a very useful quality-control tool for the importing wordnet, and this has indeed been the case for Turkish WordNet:

While building a wordnet for Turkish, the so-called “expand model” [4, p. 52] was used and synsets were constructed by providing translation equivalents for PWN synset members. Following the translation phase, a series of relations, e.g. STATE_OF relations, were imported from PWN. Since Turkish employs a morphological process to encode STATE_OF relations, the list of Turkish translation equivalents contained several morphologically-related pairs like “deli-deli**lik**” (mad-mad**ness**), “garip-garip**lik**” (weird-weird**ness**), etc. Pairs that violated this pattern probably involved mistranslations or some other problem, and the translation method provided a way to detect such mistakes.

ii. Semantically-related lexical items in the importing language are not morphologically related to each other.

In this more interesting case, the semantic relation is morphologically generated in the exporting language (Turkish) but not in the importing language (English)[3]. The causation relation between the lexical items “yıkmak” and “yık**ıl**mak”, for instance, 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. Our method thus provides a way of enriching a wordnet with semantic information imported from another wordnet. Furthermore, the proposed method improves overlap among different wordnets as they borrow semantic links from each other.

### 2.3 Monolingual and/or Multilingual Context

A possible application in a monolingual and/or multilingual context is to automate the preparation of formal and thus machine-tractable synset glosses, based on the information imported from another language’s wordnet. Equipped with the information that the Turkish synsets for “yıkmak” (tear down) and “yık**ıl**mak” (collapse) are linked to each other via a “CAUSES” relation, one can safely claim that the English synset “tear down” can be glossed as “cause to collapse”. Similarly, the builders of a Turkish wordnet can safely claim that their synset for “yıkmak” can be glossed as “yık**ıl**masına neden olmak”, which is the Turkish equivalent of “cause to collapse”.

## 3 Methodology

The methodology that will enable the above-described applications involves the following language-independent steps:

### 3.1 Determining the Derivational Affixes

All derivational affixes in the exporting language are potential candidates. Some of these have a perfectly regular and predictable semantics, while some others do not. Affixes can also be ranked according to their productivity. An affix that can be attached to almost any root in the language in question is regarded as a productive affix. Thus, two criteria have to be taken into consideration while deciding to include an affix in the list: (i) the regularity of its semantics; and (ii) its productivity.

### 3.2 Constructing Morphosemantically-Related Pairs

Using a wordlist available to the exporting language, we extract all instances containing the affix we are interested in. Simple regular expressions are sufficient for this task. We then feed all of these instances to a morphological analyzer. If there is at least one morphological analysis that suggests the expected derivation process, this instance is included in the list of potential pairs.

[3] This phenomenon has also been discussed in [5, p. 11]

The morphological analysis also provides us with the root involved in the derivation process. Thus, we obtain a list of pairs such as "teach-teach**er**" or "hang-hang**er**".

Almost all candidates which seem to, but do not actually, contain the relevant affix (such as moth-moth**er**) can be automatically eliminated by using morphological analysis results. In the case of the pair "moth-mother", the morphological analysis of "mother" does not contain the analysis "moth+Agent" and this pair can thus be safely eliminated from the list.

### 3.3 Linking the Right Synsets via the Right Relation Type

The pairs generated in the last step are merely word forms and not word senses. For the correct assignment of a semantic link, we need to assign the correct sense to both members of the pair.

Faced with the ambiguous pair "regulate-regulator" the lexicographer has to decide: (i) that the verb 'regulate' in this pair is 'regulate (sense 2)' ("bring into conformity with rules or principles or usage; impose regulations") and not 'regulate (sense 5)' ("check the emission of (sound)"); (ii) that the noun 'regulator' in this pair is 'regulator (sense 2)' ("an official responsible for control and supervision of an activity or area of public interest") and not 'regulator (sense 1)' ("any of various controls or devices for regulating or controlling fluid flow, pressure, temperature, etc.");(iii) that the resulting semantic relation involves "the second semantic effect of the suffix -or". ("the person who regulates" and not "the device that regulates").

## 4 Application of the Methodology to Turkish

Turkish, an agglutinative language with productive morphological derivation processes, employs several affixes which change the meaning of the root in a regular and predictable way [6]. There are some others which have a more complex semantics and change word meaning in more than one way. It is usually possible to specify most semantic effects of an affix and conclude, for instance, that the Turkish agentive suffix -CH[4] basically has four separate effects. Obviously, there are some fuzzy cases where it is difficult to specify the exact semantic effect. These cases usually involve semantic shifts and lexicalizations.

Table 1 illustrates Turkish suffixes we have identified as useful candidates[5].

Table 2 provides examples of morphosemantically-related pairs of Turkish words and the corresponding semantically-related pairs in English. This table clearly shows that productive and predictable morphological derivation processes in Turkish allow us to discover morphologically unrelated English words which are semantically related to each other.

The current wordlist for Turkish contains substantial numbers of words involving the suffixes listed in Table 1. We have identified the following number of instances for each suffix in Table 3

[4] Throughout the following discussion of Turkish suffixes, H represents a meta-character denoting the high vowels 'ı, i, u, ü'; A the vowels 'a, e'; D the consonants 'd, t'; and C the consonants 'c, ç'. Thus each morpheme here actually stands for a set of allomorphs.

[5] We have used the semantic relation tags defined in Princeton WordNet and EuroWordNet whenever possible. These have been indicated in boldface type throughout this paper.

**Table 1.** List of Turkish suffixes and their semantic effects (* n = noun, v = verb, a = adjective, b = adverb)

| **SUFFIX** | **POS*** | **SEMANTIC EFFECT** |
|---|---|---|
| -lAş | n-v, a-v | BECOME |
| -lAn | n-v | ACQUIRE |
| -lHk | a-n, n-n | **BE_IN_STATE** |
| -lH | n-a | 1) SOMEONE_WITH<br>2) SOMETHING_WITH<br>3) SOMEONE_FROM |
| -sHz | n-a | 1) SOMEONE_WITHOUT<br>2) SOMETHING_WITHOUT |
| -sAl | n-a | **PERTAINS_TO** |
| -(y)lA | n-b | WITH |
| -Hş | v-v | RECIPROCAL |
| -(H)l | v-v | **CAUSES** |
| -(H)t, DHr, -(H)r, -(A)r | v-v | **IS_CAUSED_BY** |
| -Hş | v-n | ACT_OF |
| -CA | a-b, n-b | MANNER |

**Table 2.** Examples of Turkish-English Pairs

| | | |
|---|---|---|
| taş | taş**laş**mak | **INVOLVED_RESULT** |
| stone | petrify | |
| iyi | iyi**leş**mek | BECOME |
| good | improve | |
| hasta | hasta**lık** | **STATE_OF** |
| sick | disease | |
| din | din**siz** | SOMEONE_WITHOUT |
| religion | infidel | |
| ölmek | öl**dür**mek | **IS_CAUSED_BY** |
| die | kill | |
| omurga | omurga**lı** | SOMEONE_WITH |
| spine | vertebrate | |

A detailed analysis of two Turkish suffixes produced the results summmarized in Table 4. The two suffixes we have investigated are -DHr and -lAş, encoding CAUSES and BECOME relations, respectively.

Despite the fact that Turkish wordnet is a small-sized resource (10.000 synsets), it contains a significant number of synsets involving these morphosemantic relations.

In only a few cases does PWN 2.0 indicate a CAUSES relation between the respective synsets. In the case of the BECOME pairs, PWN 2.0 provides the underspecified relation called "ENG DERIVATIVE".

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.

**Table 3.** Number of derived words for each Turkish suffix

| SUFFIX | # OF PAIRS | POSSIBLE RELATIONS |
|---|---|---|
| -lHk | 4,078 | **BE_IN_STATE** |
| -lH | 2,725 | WITH |
| -sHz | 1,001 | WITHOUT |
| -Hş | 991 | ACT_OF |
| -lAn | 758 | ACQUIRE |
| -lAş | 763 | BECOME |
| -DHr | 782 | **CAUSES** |
| -CA | 710 | MANNER |
| -sAl | 115 | **PERTAINS_TO** |
| TOTAL | 11,923 | |

**Table 4.** Statistics for twoTurkish suffixes

| RELATION | # IN WORDLIST | # IN TWN | # IN PWN | % OF NEW LINKS |
|---|---|---|---|---|
| CAUSES | 1511 | 80 | 18 | 77.5% |
| BECOME | 763 | 83 | 11 | 86.7% |

## 5 Conclusions and Future Work

We have tried to demonstrate that morphology offers a good starting point for enriching wordnets with semantic relations. More importantly, we have claimed that sharing morphosemantic relations across languages is an efficient way of enriching wordnets with semantic relations that are hard to discover. We have shown, at least for the case of Turkish, that there are a large number of instances involving such predictable morphological phenomena that can be fruitfully exploited for semantic relation discovery.

Future research could concentrate on automating the decision task mentioned in Sect. 3.3. The outcome of a morphological derivation process is mutually determined by the semantics of the root and the affix. Thus, there is no real "decision" involved in steps (ii) and (iii) described in Sect. 3.3. For instance, the agentive suffix -CH in Turkish is capable of producing: (i) "commodity – seller/manufacturer" pairs if the root is a marketable artefact; (ii) "person – adherent" pairs if the root is a proper noun; (iii) "instrument – musician" pairs if the root is a musical instrument, etc.

As soon as we decide that the agentive suffix -CH is attached to the root "keman" (violin) in its, say, second sense (violin as a musical instrument), we are forced to conclude that the "musician" effect OR the "seller/manufacturer" effect and NOT the "adherent" effect of the suffix is at play here. Although we cannot fully disambiguate in the absence of additional contextual and pragmatic information, we can at least rule out the possibility that the "adherent" effect might be involved.

Using the hierarchy, and more fruitfully the top ontology, of a wordnet, we can obtain additional semantic information regarding the root and predict the semantic effect the affix will have when applied to this root. The success of such a study remains to be seen.

## References

1. C. Fellbaum (ed.), WordNet: An electronic lexical database, Cambridge, MIT Press, 1998.
2. P. Vossen (ed.), EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Dordrecht, Kluwer Academic Publishers, 1998.
3. "BalkaNet: A Multilingual Semantic Network for the Balkan Languages", [online], `http://www.ceid.upatras.gr/Balkanet` (Accessed: August 23, 2003).
4. P. Vossen, "EuroWordNet General Document", [online], `http://www.illc.uva.nl/EuroWordNet/docs.html` (Accessed: August 23, 2003).
5. C. Fellbaum and G. A. Miller, "Morphosemantic Links in Wordnet", in press, Traitement Automatique de Langues.
6. K. Oflazer, "Two-level Description of Turkish Morphology", Literary and Linguistic Computing, Vol. 9, No. 2, 1994.