Defining a Framework for Semantic Categories for Turkish Nominal Morphemes

Yağmur Öztürk
Centre de recherches interdisciplinaires et transculturelles,
Université de Franche-Comté

Izabella Thomas
Centre de recherches interdisciplinaires et transculturelles,
Université de Franche-Comté

Snejana Gadjeva
Centre de recherches Europes-Eurasie,
Institut National des Langues et Civilisations Orientales

1 Introduction

As an agglutinative language, the main noun formation process of Turkish is suffixation. The roots, whether simple or complex, undergo almost no alternations during the process apart from a few exceptional ones. (1) is an example of a derived noun, with göz (en. eye) being the root to which several derivational morphemes are attached.

(1)	Turkish	English
	göz	eye
	göz-lük	eyeglasses
	göz-lük-çü	optician
	göz-lük-çü-lük	occupation of an optician

Despite this rather regular word formation process, there are very few morphological analysers processing derivational forms developed in the fields of NLP (i.e. Çöltekin, 2014), not to mention morphosemantic analysers.

In the linguistic field, several studies and language teaching books provide a description of derivational morphemes. Nevertheless, none offer a comprehensive and systematic description of these morphemes (Bazin, 1994; Adalı, 2004; Göksel & Kerslake, 2005; Korkmaz, 2014; etc) especially in regard to the description of their semantics. Some works propose a semantic study of the morphemes (Bilgin et. al, 2003; Öztürk, 2016) but as it is not the main focus of those works, they seem insufficient to be used as a base in our research.

Our research aims to systematise the semantic description of nominal morphemes in Turkish, with the goal of creating a computerised derivational resource. This line of research has been conducted in many languages within theoretical linguistics as well as the Natural Language Processing field (Talamo et al., 2016; Bagasheva, 2017; Namer et al., 2019; etc). The lack of such research in Turkish is the primary motivation behind our study.

The main question we are concerned with is how to systematise the semantic description of morphemes. How can we turn the glosses and semantic explanations, unique to each author in the studies mentioned above, into a systemic unified set of categories? A systemic description implies a depiction of, or an attempt to depict, the semantic regularities in an analysis of a derivational lexicon.

This study presents the process of establishing this set of semantic categories. First, we tried to take advantage of existing works in this field, even if none have been proposed specifically for the description of the Turkish language. In comparative semantic research in derivational morphology of Bagasheva (2017), however, Turkish is part of the languages described. As briefly discussed in section 2, this set of universal categories proposed by Bagasheva does not meet our expectations for a systemic description of the semantics of nominal morphemes. That is why we directed our research towards another resource, Wordnet (Miller et al. 1990; Fellbaum, 1998), which was not specifically designed for the description of morphosemantics, but which has been applied in recent research in French morphosemantics (Namer et al., 2019).

2 Semantic categories for affixes in Bagasheva (2017)

One of the most recent cross-linguistic research in semantics of derivational morphology is Bagasheva's work presented in Comparative semantic concepts in affixation (2017). She defines 51 semantic categories for both local (in the sense of one language), and comparative research, applied to different types of derivation.

In Bagasheva's paper, the set of semantic categories is presented in an alphabetically ordered table. The table is made of three columns: the name of the concept (Comparative semantic concept), then a short definition (Emergent meaning) of the concept and one or two examples usually in Bulgarian and English (2).

(2) a. ABILITY | Possibility to be processed in a particular way | Eng. read**able** readab**ility** / Bul. četiv**en** četiv**nost**
b. AGENT | Performer of an activity/ Name of a profession, job, title or permanent activity | Eng. kill**er** / Bul. ubi**ec** 'killer'; pek**ar** 'baker'

The proposed semantic categories are of different levels of granularity: from basic and general concepts, like ABILITY and AGENT, see e.g. (2), to more specific concepts with more detailed definitions, e.g. (3). Nevertheless, in concrete terms, the hierarchical differences between the semantic categories are neither formally established by the author, nor visible in their description.

(3) a. FEMALE | Female representative of a human type/profession | Eng. actr**ess** / Bul. čistni**ca** 'woman fastidious about cleanliness'
b. UNDERGOER | Entity that undergoes an action that changes its state | Saami čuhppojuvvot 'to be cut (of somebody)'

After a test phase on 50 items, we concluded that this set was unsuitable for the description of Turkish nominal morphemes in our framework. The test phase was divided into 4 steps (but we will not explain in detail due to limited space): 1. selection of the categories that fits or may fit in the nominal derivational process in Bagasheva's set, 2. creation of a corpus of 50 derived Turkish nouns for annotation, 3. annotation of the morphemes in the corpus, 4. verification of the annotation and observation of the results presented in Table 1.

Table 1: Results of the corpus annotation

Match to a semantic category	Number of matches	% of matches
Perfect match	25	50%
Several matches	7	14%
No match	18	36%

3 Methodology for the definition of semantic categories

From the observation mentioned in 2., we have been able to define the properties of our set of semantic categories for the description of Turkish nominal morphemes. In the following section, we will first present those. Then we will present our baseline, WordNet adapted to the description of Turkish morphosemantics.

3.1 Properties of the set

We have defined properties for the set of semantic categories, 4 in number. The following subsections present and discuss those properties.

3.1.1 An open-source project

Another side of our research is that we aim to be part of open research. This means that the finalised resources will be available for any usage in the fields of linguistics and NLP. In our study case, this research is part of a wider project and will be integrated in a morphosemantic analyser for assisted learning of derived nouns in Turkish.

This aspect of our research is an important feature to take into account in the process of establishment because it has a direct impact in the way we define our semantic categories.

Consequently, it is the main criteria in the choice of its modelisation for the resulting computerised resource we created. An ontology is by definition coherent in its structure and can be shared for different purposes. It is a modelisation of knowledge understandable by a community of people and readable by computerised resources. That is why we decided to implement the set of categories in an ontological structure we called Semantürk which aims to provide a comprehensive and structured database of semantic information for nominal derivatives in the Turkish language.

3.1.2 A hierarchically structured and organised set

A simple alphabetically ordered set is not enough to render the semantics of morphemes. Therefore, the most striking criteria is to have an organised and structured set of semantic categories. It is necessary for the semantic categories to be linked to each other in a hierarchical structure as:

- it has different levels of granularity; the hierarchy helps bring these granularities in the semantics of the morphemes out;
- moreover, it guarantees a match for all of the morphemes under study since there is a possible fallback if a morpheme does not match with a really specific semantic category.

3.1.3 Non-ambiguous categories

The most important feature is minimum ambiguity: the semantic categories have to be interpretable. The aim is to describe the semantics of the morphemes in order to comprehend the sense the morphemes convey in a derivational process. So we paid considerable attention to define the categories without ambiguity.

Another reason is related to the first property we presented in 3.1.1: this research has to be exploitable for any usage. As we previously said, in the case of our usage, the semantic categories are to be integrated in a morphosemantic analyser for Turkish learners. So the categories will be described in two aspects, first with a linguistic identifier and a definition accessible for non-linguist users.

3.1.4 Usage in other languages

To have a set of categories applicable, or at the very least partly applicable, to other languages would be an added value as future works in comparative morphosemantics would be possible. The hierarchical nature of this research would, a priori, render such a work possible since not all languages have the same levels of granularity in their morphosemantics.

3.2 Discussion: WordNet for a morphosemantic description?

Wordnet is a lexical database, applied and adapted to a large variety of languages. The semantic components present semantic primes called Unique Beginners and are 25 in number. They were not designed specifically for the description of the morphosemantics of languages as in Bagasheva (2017) but structured to form pairs of words called synsets that are semantically related. Initially built for English, WordNet has since been applied and adapted to many languages around the world and is well-established research.

A recent work Demonext (Namer et al., 2019) used Wordnet's set of semantic categories to describe French morphosemantics. It seems to offer a sufficient granularity in order to describe the morphosemantics of derived words in French. Observing morphologically related synsets, alternations in the meaning can be rendered by the attribution of Unique Beginners to each element of the pair. (4) is an example from Huguin et. al (2022) who worked on the semantic component of Demonext words. It presents the alternation mentioned for words derived with the morpheme *-eur*, alternating from ARTIFACT to PERSON as in (4).

(4) French	code; codeur {ARTIFACT; PERSON}
English	code; coder {ARTIFACT; PERSON}

Inspired by this work, we decided to test and adapt the WordNet's set for the description of derived nouns in Turkish as it meets the criteria discussed in 3.1. The semantic components of Wordnet are constructed in the hierarchical principle as well, with different levels of granularity. We adapted them to match the description of Turkish nominal morphemes.

References

Adalı, Oya. 2004. *Türkiye Türkçesinde Biçim Birimler*. Istanbul: Papatya.

Bagasheva, Alexandra. 2017. Comparative semantic concepts in affixation. In Salvador Valera Hernández & Juan Santana Lario (eds.), *Competing Patterns in English Affixation*, 33-65. Peter Lang.

Bazin, Louis. 1994. *Introduction à l'étude pratique de la langue turque*. Paris: Librairie d'Amérique et d'Orient.

Bilgin, Orhan & Kemal Oflazer. 2004. Morphosemantic Relations In and Across Wordnets. In Petr Sojka, Karel Pala, Christiane Fellbaum & Piek Vossen (eds.), *Proceedings of the Second International WordNet Conference (GWC 2004)*, 72-78. Brno, Czech Republic: Masaryk University.

Çöltekin, Çağrı. 2014. A set of open source tools for Turkish Natural Language Processing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, *Database. Language, Speech, and Communication*. Cambridge, London: The MIT Press.

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar (Routledge Comprehensive Grammars)*. London: Routledge.

Huguin, Mathilde, Lucie Barque, Pauline Haas, Fiammetta Namer, Delphine Tribout. 2022. *Guide d'annotation Demonext : Typage lexical des noms du français*. https://hal.science/hal-03638962

Korkmaz, Zeynep. 2014. *Türkiye Türkçesi Grameri Şekil Bilgisi*. Türkiye: Türk Dil Kurumu editions.

Miller, George. A., Richard Beckwith, Christiane Fellbaum, Derek Gross, & Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3:4. 235–244.

Namer, Fiammetta, Lucie Barque, Olivier Bonami, Pauline Haas, Nabil Hathout & Delphine Tribout. 2019. Demonette2 - Une base de données dérivationnelles du français à grande échelle : premiers résultats. In Emmanuel Morin, Sophie Rosset & Pierre Zweigenbaum (eds.), *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, 233-243. Toulouse, France: ATALA.

Öztürk, Seda. 2016. *Création et reconnaissance de néologismes par méthode de suffixation*. Master's thesis, Université de Franche-Comté.

Talamo, Luigi, Chiara Celata & Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure* 9:1. 72-102. doi:10.3366/word.2016.0087.