Morphological data extraction from a corpus of nouns in Turkish

Yağmur Öztürk
Centre de Recherches Interdisciplinaires et Transculturelles (CRIT),
University of Bourgogne Franche-Comté, France
yagmur.ozturk@edu.univ-fcomte.fr

In the field of computational linguistics in Turkish, very few studies are conducted in derivational morphology, especially regarding nominal derivation. Most of the existing open-source studies are focused on inflectional morphology and deal with a very small number of the most productive derivational morphemes. As can be seen in Zemberek (Akın & Akın, 2007) and TRmorph (Çöltekin, 2014), two open-source morphological analysers, the main focus is given to inflectional morphemes.

In this paper, we will present the creation of a corpus of nouns, a resource created in order to study and extract relevant data concerning derivational morphemes in Turkish. These morphemes correspond to 41 nominal suffixes that attach to nouns, previously selected from (Göksel & Kerslake, 2005) but are not yet a fixed set.

The corpus of nouns is built from the extraction of all of the nominal entries from an online Turkish dictionary (Türk Dil Kurumu Sözlükleri, *the dictionaries of the Turkish Language Association*). We plan to compare and add to it nominal entries from another online Turkish dictionary (Dil Derneği, *The Language Association*). As of now the corpus is composed of around 39 000 words of nominal morphosyntactic category.

This corpus will be used to observe and extract data around the 41 derivational morphemes. We aim to calculate their productivity and extract information regarding their distribution based on the behavior of each morpheme in the corpus.

This research is part of a wider project which aims to provide a set of resources in Turkish to overcome the lack of existing open-source resources in the fields of derivational morphology in Turkish NLP. The final purpose of the creation of these resources is to build a complete morphosemantic inventory of Turkish noun to noun derivational morphemes to explicit word-meaning construction patterns in Turkish through equivalences in French word-meaning construction patterns.

References

AKIN A.A. & AKIN M.D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10, 1-5.

ÇÖLTEKIN Ç. (2014). A Set of Open Source Tools for Turkish Natural Language Processing. In CALZOLARI N. et al. Ed., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 1079–1086.

Dil Derneği, (*Language Association*). Accessed 14 March 2022, http://www.dildernegi.org.tr/TR,451/turkce-sozluk.html.

GÖKSEL A. & KERSLAKE C. (2005). *Turkish: A Comprehensive Grammar*. Routledge Comprehensive Grammars. London: Routledge.

Türk Dil Kurumu Sözlükleri (*the dictionaries of the Turkish Language Association*). Accessed 14 March 2022, https://sozluk.gov.tr/.