Available online at www.sciencedirect.com

ScienceDirect

Lingua xxx (2017) xxx–xxx

Lingua

www.elsevier.com/locate/lingua

A computational morphological lexicon for Turkish: TrLex

Ozkan Aslan [a,*], Serkan Gunal [a], B. Taner Dincer [b]

[a] Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye
[b] Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, Turkiye

Received 5 May 2017; received in revised form 15 December 2017; accepted 16 January 2018

Abstract

A morphological lexicon that is a computational source should be considered together with derivational morphology especially for agglutinative languages. To the best of our knowledge, in the Turkish language there has been no study that analyzes the derivational suffixes on the lexicon in a computational paradigm. This study provides a very rich lexical resource, filling a gap in the field, and would hopefully lead to new related studies as well. The morphological lexicon can be used in morphological analysis as well as in several other tasks, such as stemming and part of speech (POS) tagging. In this study, we introduce a morphological lexicon named TrLex and present its components, preparation processes and some statistics. We observed that more than half of the single-word lemmas (56.7%) are in the derived structure. Since the word formation in Turkish prefer the morphological processes, this number is higher than the rate of compound-type words (2.7%). As a result of the work, we obtained a knowledge-intensive data table including several fields such as form, structure, semantic information. We also extracted Lexical Markup Framework (LMF) formatted file containing only morphological and POS information and made the file freely available.

Keywords: Morphological lexicon; Morphology; Derivation; Compounding; Turkish

1. Introduction

Natural language processing (NLP) is a wide field of study to which many disciplines and researchers contribute. Some of the studies in this area use rule-based methods which need structured linguistic data. Within the frame of hybrid approaches or with the aim of comparison, statistical methods may also apply structured data as in rule-based methods. The lexicons are typical examples of structured linguistic data. Lexicon studies have been performed in many fields (Zweigenbaum et al., 2005; Lagos et al., 2015; Lewin, 2016), for many languages (Yip, 2000; Oflazer and Inkelas, 2003; Yap et al., 2010; Heyd, 2015) and for various periods (Dalby, 1965; Silvestre, 1998; Stoll, 2015). Dictionaries are a little bit different from lexicons: while dictionary refers mostly to a physical object that provides information for human users, lexicon refers to the entries used in computer programs (Hayashi and Ishida, 2006). Moreover, dictionaries are subject to

* Corresponding author at: Department of Computer Engineering, Anadolu University, 26555 Eskisehir, Turkey
E-mail addresses: ozkanaslan@anadolu.edu.tr (O. Aslan), serkangunal@anadolu.edu.tr (S. Gunal), dtaner@mu.edu.tr (B.T. Dincer).

https://doi.org/10.1016/j.lingua.2018.01.003
0024-3841/

some restrictions for practical purposes, such as listing the words in alphabetical order, but lexicons can be excluded from this constraint (Ježek, 2016).

Lexicons can be divided into various different categories according to their types, scopes and purposes. Lexicons used in NLP are computational lexicons. Computational lexicons are simply manipulable and machine-readable versions of usual dictionaries (Litkowski, 2005). They can provide information in various formats such as Extensible Markup Language (XML), a network, or a database. Basically, in their itemized context, computational lexicons contain data under various categories, such as phonetic, morphological, syntactical, grammar, semantic, etymological and ontological. Thus, they are used in many fields, such as word-sense disambiguation, information extraction, question answering, text summarization, and speech recognition (Litkowski, 2005).

Several studies on computational lexicon exist in various languages: Comlex Syntax (Grishman et al., 1994) is a lexicon that consists of 38,000 English headwords and syntactic (subcategorization) information. CLIPS (Ruimy et al., 2002) is a multi-layered Italian computational lexicon that consists of 55,000 lemmas that are coded phonologically, morphologically, and syntactically. Maltilex (Rosner et al., 1998) is a lexicon that includes morphological and syntactic information for Maltese. PDT-Vallex (Urešová, 2009) is a Czech lexicon that includes valency information obtained from annotating in the Prague Dependency Treebank (PDT). Lefff (Sagot, 2010) is an extensive, freely available French lexicon that consists of morphological and syntactic information. It is developed within the Alexina framework, which is compatible with the Lexical Markup Framework (LMF). CML (Tadić and Fulgosi, 2003) is a morphological lexicon that is produced for Croatian and it includes two sub-lexicons, one derivative and the other inflectional. SKEL (Petasis et al., 2001) is a morphological lexicon that is prepared in Greek. Leffe (Molinero et al., 2009) is an extensive morphological and syntactic lexicon prepared for the Spanish language.

A morphological lexicon which is a computational source should be considered together with derivational morphology especially for agglutinative languages. To the best of our knowledge, in the Turkish language, there has been no study that comprehensively analyzes the derivational suffixes on the lexicon in a computational paradigm. This study will primarily provide a very rich lexical resource, filling a gap in this field, and will hopefully also lead to new related studies. The morphological lexicon can directly be used, especially in morphological analysis, which is one of the major tasks of NLP. In addition, it may also be used in several other processes such as stemming and part-of-speech tagging. Furthermore, the lexical variety of derivational suffixes can be analyzed by specifying base-suffix pairs. Such a lexicon can provide numerous analyses for the relations between forms, structures, and meanings. In this sense, the lexicon will function as a very rich resource.

In this study, we introduced a morphological lexicon named *TrLex*[1] and presented its components, preparation processes and some statistics. In the paper, first, we briefly explained the morphological structure of Turkish. Following that, we described the content of the data, morphological segmentation and phonological tagging processes. Lastly, we explained the structure of LMF in the Material and Methods. In the Results, we first presented some general statistics of the lexicon. Later, we presented the statistics referring to the results of morphological segmentation and phonological tagging in figures and tables. In the Discussion, we discussed and interpreted the findings that are presented in Results. Finally, in the Conclusion and Future Work, we presented the conclusions of this study and expressed our predictions related to further research.

2. Turkish morphology

Turkish is an agglutinative, head-final and free-constituent-order language. The sentence structure in Turkish is generally in subject–object–predicate order. It has one of the richest suffix systems in terms of morphological properties. Turkish is a challenging language for computational linguistics studies since it is a free-order language and contains many suffixes.

Phonetic change rules, intertwined with morphological processes, are largely regular. However, there are many cases that are in contrast with its regularity-based property of being a language that is ‘‘pronounced as written’’ (Altun, 2010). Both items from foreign languages and its inner units can adversely affect the formal regularity of Turkish. A comprehensive study on the phonetic disharmony of Turkish was performed by Clements and Sezer (1982). Phonetic change causes a variation in the allomorphs of the suffixes. In this variation, the significative rule depends on palatal and labial harmony of the last vowel of the base (Deny, 1955).

Although the word formation and the inflection are performed with the help of suffixes in Turkish, there are some items that are especially used for terms and seem like prefixes, such as *öz- (self)*, *iç- (inner)*, *dış- (outer)*, *ilk- (first)*, and *ön- (pre)*. Nevertheless, they cannot be regarded as prefixes as they are also meaningful on their own (Şahin, 2006). The suffixes of

[1] This lexicon was used in a syntactic disambiguation system (Aslan, 2017).

Please cite this article in press as: Aslan, O., et al., A computational morphological lexicon for Turkish: TrLex. Lingua (2018), https://doi.org/10.1016/j.lingua.2018.01.003