Storage and Computation of Multimorphemic Words in Turkish

Rabia Ergin¹, Timothy O’Donnell² & Emily Morgan³

¹Cognitive Science Program, Boğaziçi University; ²Department of Linguistics, McGill University; ³Department of Linguistics, University of California, Davis

Introduction

Background

- Decompositional theories suggest that morphologically complex words are stored in terms of their component morphemes (e.g., Marslen- Wilson & Zhou, 1999; Taft & Forster, 1975).
- At the other end of the spectrum are full-listing theories, which suggest that morphologically complex words are stored as unanalyzed wholes in the mental lexicon (e.g., Butterworth, 1983; Manels & Tharp, 1977; Seidenberg & Gomez, 2000).
- Between these two extremes are dual-route, dual-mechanism, or dual representation theories. Under these accounts, a word may be processed using one or both routes, depending on various properties such as whether it is a word or non-word, high or low frequency, derived or inflected, and regular or irregular (e.g., Baayen, Dijkstra & Schreuder, 1997; Pinker, 1999; Jackendoff, 2002).
- To date, however, the vast majority of the research on this topic has come from Indo-European languages that contain relatively simple morphological systems.

Present Investigation

- **Goal:** In the present study, we apply a probabilistic tradeoff-based model of morphological storage and computation—known as fragment grammars (O’Donnell, 2015)—to make predictions about which combinations of Turkish morphemes might be stored and which computed.
- **The tradeoff-based approach** is designed to distinguish freely combining productive units in a language (e.g., words and morphemes) from recurring patterns which do not generalize but are rather stored together within larger structures. This does it by optimizing a balance between two competing biases.
- The **first bias** favors smaller more compact lexicons with highly reusable units. The **second** favors simple derivations of individual forms, involving fewer lexical items.
- These two biases are opposing—if units are smaller in general, the lexicon will contain few items, but more units will be needed to derive individual forms. On the other hand, if units are larger then forms can be derived using fewer steps, but the lexicon will have to contain many, less reusable forms.

Turkish

- Turkish is an Altaic language that has an extremely productive morphological system, realized primarily through suffixation. The following complex words taken from the METU Turkish Corpus (Atalay, Öflazer, & Say, 2003) exemplify the morphological system:

  - (1) Göz–bu
  - Eye—DER

- The “one with the glasses”

- (2) Güven-m–i-yer–du – m

- “I was not able to trust.”

- Turkish morphology is highly productive—Hankamer (1989) estimates that the average multimorphemic word contains 4.8 morphemes and that each verb can have over 2,000 inflectional forms. Considering the huge number of entries and the highly regular nature of Turkish morphology, it has been claimed that Turkish speakers must rely on decomposition.

- Several experimental studies present suggestive evidence for a dual-route architecture (e.g., Gürel, 1999; Ergin, Jackendoff & Cohen-Goldberg, 2014).

Methods & Evaluation

**Over the 500 combinations, we were able to categorize 421 out of the patterns listed in Table 1.**

**Nouns:**

- Of the 500 combinations, we were able to categorize 421 out of the patterns listed in Table 1.

**Adjectives:**

- 406 out of 500 combinations are categorized into patterns listed in Table 2.

**Conclusions:**

We have reported the results of a preliminary study deriving predictions for storage and computation of word forms in Turkish—a morphologically rich language. Despite the fact that Turkish is highly regular, we found that the probabilistic tradeoff-based model of O’Donnell (2015) predicted a number of patterns of nouns and adjectives which were plausible candidates for storage.

To evaluate the Fragment Grammar, we examined the probable nominal and archetypical direct tree fragments (500 of each) that contained at least two leaves at the frontier of the fragment.

**Figure 2.** Two tree fragments discovered by our model. On the right hand side is a tree fragment which predicts that the stored combination of morphemes marking the third person possessive and ablative case (“by his/her N”) can freely combine with an arbitrary root variable of category Noun. On the left hand side is a tree fragment which stores all surface morphemes down to leaves (tartan/”by”).

By optimizing this tradeoff on a particular input dataset, the model makes specific predictions about the pattern of computation and reuse in a language. The tradeoff-based approach has been used successfully in a number of morphological systems, but has not been applied to a morphologically rich agglutinative language such as Turkish.

References


Corresponding Author: rabia.ergin@boun.edu.tr