Storage and Computation of Multimorphemic Words in Turkish

Rabia Ergin[1], Timothy O'Donnell[2] & Emily Morgan[3]

[1]Cognitive Science Program, Boğaziçi University; [2]Department of Linguistics, McGill University; [3]Department 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; Manelis & Tharp, 1977; Seidenberg & Gomerman, 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.

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, Oflazer, & Say, 2003) exemplify the morphological system:

  (1) Göz–lük –lü
  Eye–DER –DER
  "The one with the glasses"
  (2) Güven–e – m – iyor– du – m
  Trust – ABIL–NEG – PROG – PAST – 1stSg.
  "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).

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. It does this 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.

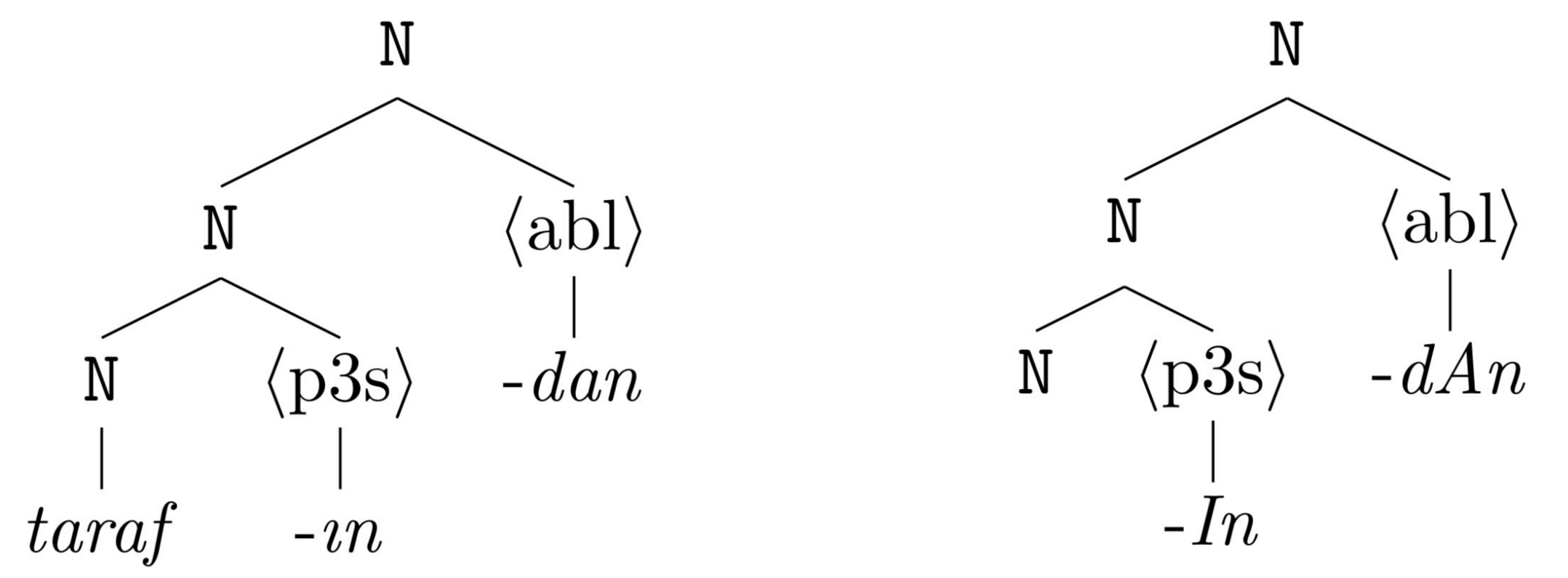

Figure 1. Two tree fragments discovered by our model. On the right hand side is a tree fragment which predicts that the stored combination of morphemes marking the third person possessive and ablative case ("by his/her N") can freely combine with an arbitrary root variable of category Noun. On the left hand side is a tree fragment which stores all surface morphemes down to leaves (tarafından "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.

Methods & Evaluation

In this study, we train the model on a corpus of Turkish words. The model predicts specific stored fragments of structure (together with their probabilities); we discuss these predictions.

- **Corpus construction** We begin with the METU-Sabancı Turkish Treebank (Atalay, Oflazer, & Say, 2003), a corpus of 173,469 Turkish word types. We obtain initial morphological parses for all words in the corpus using the TRmorph morphological analyzer and disambiguator (Çöltekin, 2014). For the majority of word types in the corpus, this disambiguator assigns a single best morphological analysis, which we take as the single analysis for our corpus. For 33,152 word types, TRmorph cannot assign a best analysis. For these words, we hand-constructed a set of heuristics to narrow the list of possible analyses to only those that were truly possible in Turkish. 54,624 total analyses for 22,963 word types passed our heuristic filters—that is on average, each of these word types had 2-3 possible analyses in our corpus. We included all such analyses in our corpus, distributing token counts obtained from the METU corpus evenly over word types.
- **Fragment Grammar analysis** We ran the fragment grammars stochastic search algorithm through the corpus for 1000 sweeps. Parameters were set to Pitman-Yor discount parameter was set to 0, the Pitman-Yor concentration parameter was set to 1, and all pseudocounts were set to 1 for all outcomes on all Dirichlet-Multinomial distributions (see, O'Donnell, 2015 for details on the meaning of parameter values). We performed all following analyses on the best scoring grammar output by the model.
- To evaluate the Fragment Grammar, we examined the probable predicted nominal and adjectival tree fragments (500 of each) that contained at least two leaves at the frontier of the fragment.

Nouns

Of the 500 combinations, we were able to categorize 482 into one of the patterns listed in Table 1.

Ex: N = N + POSS – a nominal root followed by a possessive maker is the most common combination and all of the combinations having this pattern involve a specific root.

(3) Eğitim Bakanlığ – ı
Education Ministry – POSS "
Ministry of Education"

Adjectives

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

Ex: Adj = V + <part:pres> is the most frequent combination in the adjective list.

(4) Adj = yap <part:pres>
yap–an
"the one who is doing"

Table 1. Noun combinations

	# of combinations	
Pattern	with specific root	with generic root
1. N = N + POSS	112	0
2. N = N + POSS + CASE	69	17
3. N = N + CASE	85	0
4. N = V + INF	39	0
5. N = N + PL	23	0
6. N = N + PL + POSS	20	3
7. N = V + vn:past/fut + POSS + CASE	11	6
8. N = N + PL + CASE	8	6
9. N = V + INF + POSS	10	2
10. N = N + PL + POSS + CASE	2	8
11. N = V + INF + CASE	7	3
12. N = V + TENSE + POSS	4	2
13. N = N + cpl-past/pres /evid/cond + PERS	0	6
14. N = V + vn:past/fut + PL + POSS + CASE	1	4
15. N = N + LIK	5	0
16. N = V + <yis>	5	0
17. N = N + LIK + POSS	2	2
18. N = V + INF + PL	3	0
19. N = N + CI	2	0
20. N = N + CI + LIK	1	1
21. N = Other	5	8

Table 2. Adjective combinations

	# of combinations	
Pattern	with specific root	with generic root
1. Adj = V + <part:pres>	118	0
2. Adj = N + LI	96	0
3. Adj = V+ part:past/fut+POSS	79	7
4. Adj = N + SAL	18	0
5. Adj = N + POSS + LOC + KI	17	1
6. Adj = N + SIZ	17	0
7. Adj = N + LIK	9	0
8. Adj = Adv + KI	9	0
9. Adj = V/N+lAn+part:press/fut	9	0
10. Adj = N + LOC + KI	6	1
11. Adj = V <neg> <part:pres>	3	1
12. Adj = Adj + lAs + part:pres	1	1
13. Adj = Other	9	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. We highlighted the comparison between patterns which contained a stored root and those which contain a generic root variable. These particular cases can form the basis for experimental manipulations testing the psychological reality of our claims.

References

① Atalay, N. B., Oflazer, K., & Say, B. (2003). The Annotation Process in the Turkish Treebank. Presented at the Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC, Budapest, Hungary.

② Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: evidence for a parallel dual- route model. Journal of Memory and Language, 37, 94– 117

③ Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production* (Vol. 2). New York: Academic Press.

④ Çöltekin, Ç. (2014). A Set of Open Source Tools for Turkish Natural Language Processing (pp. 1079–1086). Presented at the In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).

⑤ Gürel, A. (1999). Decomposition: To what extent? The case of Turkish. *Brain and language*, *68*(1-2), 218-224.

⑥ Ergin, R., Jackendoff, R., Goldberg-Cohen, A. (September, 2014). The processing of multimorphemic words in Turkish: Evidence for a dual-route. Poster presented at 8th International Conference on the Mental Lexicon, Montreal, CA.

⑦ Hankamer, J. (1989). Morphological parsing and the lexicon. In *Lexical representation and process* (pp. 392-408). MIT Press.

⑧ Jackendoff, R. S. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

⑨ Manelis, L., & Tharp, D. A. (1977). The processing of affixed words. *Memory and Cognition,* 5, 690-695.

⑩ Marslen-Wilson, W., & Zhou, X. (1999). Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes*, 14, 321-352.

⑪ O'Donnell, T. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage.* MIT Press.

⑫ Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. New York: Perennial.

⑬ Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, 4, 353-361.

⑭ Taft, M., & Forster, K.I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14, 638-647.

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