Morphological Parsing, the Mental Lexicon and the Dictionary

Petek KURTBÖKE, Melbourne, Australia

Abstract

This paper deals with the 'word' status of suffixes in a Turkish speaker's mental lexicon and their perception as units. Listing of individual suffixes and parsing methods based on isolation techniques are criticized. The computation of various suffix strings in Turkish leads to the early conclusion that suffixes, which tend to co-occur frequently, lose their individual meanings and acquire unit status. A treatment of these as units may remove certain delimitations in computational applications and the ambiguity problem frequently encountered when parsing Turkish texts.

1 Introduction

This paper has a few starting points all coming from different directions. In chronological order, these can be summarised as follows: an early work on the frequency counts of Turkish affixes [Pierce 1961, Pierce 1962], a seminal article on the word status of Turkish suffixes in the mental lexicon [Hankamer 1989], a proposal for a new dictionary of Swahili, a language with rich affixational morphology [Bwenge 1989], corpus-driven work on collocation and delexicalisation [Sinclair 1991], and an ongoing Turkish-English dictionary project [Kurtböke/Beadle *in progress*]. I will first deal with each work seperately and then attempt to link the relevant points.

1.1 Frequency counts of affixes

In a pioneering count of the frequencies of Turkish affixes conducted in Turkey between 1957 and 1960, Pierce listed 110 frequent affixes occurring in spoken and 139 frequent affixes occurring in written Turkish. The rationale for the calculation was to break up each word [i.e. a stem plus all of its affixes] and count each morpheme in a corpus of Turkish texts. öne of the most striking features of Turkish", Pierce concluded, was "the fact that in approximately 140,000 words of running text, 112,001 affixes were found" i.e."for every five stems, four affixes occurred in the text" [Pierce 1961]. He also recorded whether the most frequent affixes were *derivational* and *inflectional* and their relative percentage in the total frequency for all affixes. He found that out of the twenty most frequently found affixes only four were derivational. Pierce also acknowledged, however, that

> [j]ust what these figures mean is difficult to say, except that it is extremely interesting to see that only twenty morphemes accounted for such a large percentage of the total text studied, and that only 36 inflectional affixes accounted for nearly 74 percent of the total [Pierce 1962, 41].

1.2 Word recognition and parsing

The claim that at least in the mental lexicon of the speakers of agglutinative languages, word recognition and parsing operate differently from that of the speakers of European languages [Hankamer 1989], has not been investigated further. The starting point of Hankamer's reasoning was that agglutinative languages with complex morphology such as Turkish allow words of indefinite length produced by means of iterative loops. Two significant observations were that these forms are very common and recognized as words by native speakers of Turkish. The main premise of Hankamer's parsing model was the recognition of the root before affix stripping took place. Morphological parsing in word recognition, Hankamer argued, should proceed from left to right, at least for agglutinative languages with rich suffixation. The reason for this strategy is that "the left-to-right recognition approach narrows the choice of possible suffixes at every step to suffixes that can combine with a stem of the current stem category" [Hankamer 1989, 402]. Even when narrowed down, still a very high number of forms can be associated with any verb or noun root, which would go far beyond the storage capacity of the brain. Therefore, Hankamer concluded that "the human mind has, or is capable of acquiring, a parsing mechanism that allows the recognition and understanding of words of impressive complexity" [Hankamer 1989, 404].

1.3 Affixational morphology and the dictionary

[Bwenge 1989] argued that existing dictionaries of Swahili, which has agglutinative structure like Turkish, suffer from the traditional arrangement of nominal and verbal items. He complained that all-existing Swahili dictionaries "have been characterized by inadequacy, inconsistency and lack of systemacity in their handling of the phenomena of affixational morphology" [Bwenge 1989, 7]. He then proposed a new dictionary design for Swahili with a clear and systematic presentation of affixes which would be "treated as a property of the lexicon" [Bwenge 1989, 16].

1.4 Collocation and delexicalisation

A much-debated question, more enthusiastically over the last decade, has been *what constitutes a lexical unit*? The notion of collocation has been elaborated in response to this question, and its lexicographical and grammatical aspects have been investigated e.g. [Sinclair 1991]. Combined approaches have also been available. Previously, [Pawley/Snyder 1983] drew attention to *prefabricated patterns*, which "form a high proportion of the fluent stretches of speech heard in everyday conversation". Then, [Nattinger/DeCarrico 1992] proposed a *lexico-grammatical unit* to be used as the central unit in teaching. They placed the *lexical phrase*, described as 'chunks' of language of varying length "somewhere between the traditional poles of lexicon and syntax". [Kjellmer 1991], in a combination of corpus and psycholinguistic approaches, has come up with a typology granting that the borderlines were fuzzy. The categories were *Fossilized phrases*, *Semi-fossilized phrases* and *variable phrases*.

Descriptive parameters of collocation have been problematic in machine translation, as well. For example, [Storrer/Schwall 1995] proposed a combination of corpus and dictionary-based methods for the analysis, generation and transfer of *multiword lexemes*. They stressed two properties

of multiword lexemes: their status in the mental lexicon and their complex nature in contrast with simplex words, as well as their aspects of non-compositionality and non-substitutability.

In spite of extensive work on collocations over the past few years, however, the concept is still fuzzy:

> But perhaps this is exactly as it should be, given the nature of the language: some combinations will be more fixed than others because some concepts are better established than others. [...] As a heuristic method, statistics will be useful to solve this problem of where collocations stop and free combinations begin [Van Der Meer 1998, 319].

A useful notion explored in relation to the notion of collocation has been *delexicalization* defined as a reduction of the distinctive contribution made by a word to the meaning [Sinclair 1991:113]. Delexicalisation has been central to the study of collocation as delexical words act in conjunction with other words and share their meaning [Sinclair and Renouf 1988, Partington 1993]. The reduction of delexical verbs to suffix status has also been demonstrated [Kurtböke 1998a].

1.5 A new Turkish-English dictionary

When the project started four years ago[1], the original idea was to create an experimental computer-based dictionary from a corpus. At the earliest stage, the obvious thing to do seemed some improvement to the existing Turkish-English dictionaries, as is usually done. *Ozturk corpus* [Kurtböke 1998b] provided the textual database from which the wordlist and concordances were generated. The analysis of these would place the word forms into the lexical database from which the dictionary would be produced. However, in the course of the analysis of the word forms, an average verb appeared to have over 250 forms, realized through suffixation. While it was not a morphological revelation nobody had experienced before, a corpus-driven analysis as well as an interpretation of the form, function and meaning of these suffixes has been lacking. This search for an improved computational analysis and parsing of Turkish suffixes paves the way for a consideration of their status in the mental lexicon and their place in the grammar.

2 Word status of suffixes

The point to be emphasized in Pierce's study [Pierce 1961, Pierce 1962] is that suffixes make up a very considerable portion of a Turkish text, and some twenty of these suffixes are very frequent. This confirms that suffixation is the driving force of the language and requires more attention. However, any attempt to replicate Pierce's analysis of individual Turkish affixes in computational terms will bring little joy to the researcher and probably be abandoned halfway. In fact, Pierce's counts were carried out manually as he had little computational assistance at the time. Had Pierce analysed Turkish suffixes computationally, he would have experienced the difficulty of extracting from the corpus such forms as, for example, the accusative [i] which,

in accordance with the vowel harmony rule, may also appear as [1], [u] or [ü]. Any restrictions imposed on the search would still fail to yield the result desired due to some other words with similar endings. The problem is less evident as we move on from one-letter suffixes to longer ones, although ambiguity remains a problem. The conventional approach to morphological analysis is to break up all the suffixes and identify each suffix separately as Pierce had done. This is due to the somewhat inferior status of suffixes seen as small particles without much independence of their own. However, Bwenge's research on Swahili indicates that they should be treated as a property of the lexicon. This brings us to their place in a Turkish speaker's mind.

Hankamer's work, on the basis of his computation of Turkish suffixes, suggested that human mind is capable of parsing words of impressive complexity. However, while in theory a Turkish speaker may produce words of indefinite length with all the available suffixes, this does not happen in reality. The longest stretch of suffixes on an average word in the corpus contains 6 to 7 suffixes. This seems to be the point where previous thinking on Turkish suffixes requires refinement. That is, a shif of focus is required from the the probable length of suffix-stretches and permutations, which have so far attracted more attention than their other aspects such as combinability and co-occurrence, to their unit status.

If suffixes in Turkish have the word status in a Turkish speaker's mind, then these suffixes are subjected to the processes of collocation and fossilization to the same degree as words. They are restricted in terms of position and some suffixes attract while others repel each other. Where strong combinations and positional preferences stop and free combinations begin can only be determined on the basis of the frequencies and concordances. However, it is clearly unhelpful to break up these suffix clusters, as it is to break up word collocations, and try to assign a meaning or a function to each form. Corpus analysis shows that a considerable number of combinations is possible, however, some of these combinations are very frequent and seem to have acquired unit status. In some very frequent suffix combinations, similar to frequent word collocations in the language, there appears to be a loss of meaning and function. Unfortunately, in the existing Turkish dictionaries suffixes are not covered with respect to their combinability and positional preferences.

2.1 Examples

The tradition so far has been to identify each suffix, label it with an appropriate tag and parse the text. However, corpus data[2] show that suffix combinations, at least certain formations, have a higher frequency than the individual suffixes as in for example the forms of the verb *bil-* [know]:

Root	suffix	frequency
Bil	dir 1	
Bil	dirdi	21
Bil	dirildi	42

Over 1000 such combinations [root+suffixes] in Turkish occur on a regular basis [Beadle 2000] challenging the view that a Turkish speaker reshuffles an endless number of inflectional suffixes every time a new conjugation is required in speech. An

Ordinary Turkish verb stem accomodates between 200-400 suffix strings. At the highest extreme there is the verb ol-[be], which takes on twice as many strings as any other verb. While it is impossible to illustrate all of these forms here, let us take a look at some 63 strings with frequencies varying between 10-988 in a list of ol- forms:

1.	V-abilirsiniz 000010
2.	V-mayı 000010
3.	V-ursunuz 000010
4.	V-amaz 000011
5.	V-malarını 000011
6.	V-masa 000011
7.	V-masının 000011
8.	V-mayabilir 000011
9.	V-mazsa 000011
10.	V-saydı 000011
11.	V-uşturduğu 000011
12.	V-abildiğince 000012
13.	V-anlara 000012
14.	V-duğunun 000012
15.	V-madığına 000013
16.	V-mayacak 000013
17.	V-unuz 000013
18.	V-makla 000014
19.	V-maması 000014
20.	V-unması 000014
21.	V-urlar 000014
22.	V-abiliriz 000017
23.	V-malı 000017
24.	V-duğumuz 000018
25.	V-duğundan 000018
26.	V-duğunda 000019
27.	V-muştur 000020
28.	V-maktadır 000022
29.	V-maktan 000022
30.	V-manın 000022
31.	V-maz 000022
32.	V-duklarını 000023
33.	V-abilirler 000024
34.	V-acağı 000025
35.	V-masını 000027
36.	V-madan 000030
37.	V-unur 000031
38.	V-anların 000032
39.	V-abilecek 000033
40.	V-malıdır 000033
41.	V-manız 000035
42.	V-uyor 000035
43.	V-acağını 000036
44.	V-ayları 000037
45.	V-maya 000037
46.	V-dukları 000038
47.	V-maları 000038
48.	V-uşan 000040
49.	V-duğunuz 000041
50.	V-madığı 000041
51.	V-dukça 000042
52.	V-masına 000042
53.	V-anlar 000045
54.	V-duğuna 000045
55.	V-madığını 000046
56.	V-ursa 000046
57.	V-acaktır 000103
58.	V-abilir 000121
59.	V-mayan 000138
60.	V-ması 000175
61.	V-duğunu 000285
62.	V-duğu 000320
63.	V-arak 000988

One way of looking at the permutations is to analyze these forms as a property of *ol-* and investigate them as individual cases. This may be the preferred approach for the lexicographer as the concern remains the semantic representation of the lexical unit. A more productive way, however, is to see whether these forms are distributed across verb stems as well. They do[3]. The task here is to compute each of these strings as full words as they attach to numerous verb stems. The view so far has been that a string like /-mayabilir/ [‘may not be able to’] can be broken into ma+[y]abil+ir [#8 above] with each suffix carrying a meaning and a function. Accordingly,

within this framework, ‘mayabilir’ is made up of three slots waiting to be filled up:

```
Verb + slot1 + slot2 + slot3.
```

On the contrary, ‘mayabilir’ is a fossilized string, which easily attaches to other stems. That it is fossilized is clearly demonstrated by the epenthetic ‘y’. The same argument may be put forward for each of the combinations listed in the table.

3 The proposal and Conclusion

It follows from the foregoing section this way of looking at agglutinative morphology has implications for the studies of the mental lexicon and NLP applications. An example of agglutinative languages is Turkish, which has been frequently described by western scholars [e.g. Deny, Godel, Németh, Lewis etc] in line with western grammars. Descriptions by Turkish scholars have also been under the western influence [e.g. Gencan]. Later grammars [e.g. Underhill 1976], instead, described Turkish in terms of the theoretical models developed elsewhere. A more recent trend within the Natural Language Processing approach to Turkish is to test theoretical models such as Lexical-Functional grammar, Systemic-Functionalist theory etc. on Turkish. Although these models have a universal claim, they are basically grammars of English. According to [Quirk 1992] there remains a certain tension between "[a] those who want to know as much as possible about language [...] and [b] those who want to know as much as possible about what the computer can do".

While both approaches are equally valid and "potentially complementary", they cannot be seen as two ways of achieving the same goal. "How we choose between them, or where we put our emphasis, will affect public policy, the support of funding, and hence the direction and future of our research.". The situation summarised by Quirk applies to the current corpus research on Turkish, which remains along the lines of computer engineering rather than corpus-driven lexical research. However, parsing models based on western languages do not clearly work in the description of agglutinative languages such as Turkish since the speakers recognize suffix strings as words. The probability of suffix co-occurrences and the parsing of these strings should be based on corpus-driven knowledge of affixational morphology rather than knowledge imported from languages with limited suffixation mechanisms. Previously computed frequencies of individual suffixes in a corpus of Turkish revealed that some 75 percent of the text is made up of inflectional suffixes. This observation contradicts the computational approaches to affixational morphology concerned mainly with probabilities of permutations. Studies of suffix strings assume up to millions of suffixes can occur together without disturbing the pattern and the speaker would recognize this even if this pattern has never occurred before in speech [Hankamer 1989]. While in theory these suffixes can recur a million times and in various combinations, in everyday speech they don’t. Instead, they regularly occur in certain patterns. The present proposal, therefore, criticizes the slot-and-filler approach to morphotactics, and promotes the unit status of suffixes and the patterns in which they occur.

Notes

[1] The initiation of the project became possible through a Euralex Verbatim Award granted to Petek Kurtböke in 1996.

[2] Ozturk Corpus [Kurtböke 1998a]

[3] See [Kurtböke 1998b] for a similar treatment of *yap-* [do].

References

[Beadle 2000] Beadle, J (2000) *A Survey of existing SIL software for the construction of a Turkish Lexical Database.* Summer Institute of Linguists.

[Bwenge 1989] Bwenge, C (1989) Lexicographical treatment of affixational morphology: a case study of four Swahili dictionaries. In G James [ed] *Lexicographers and Their Works.* University of Exeter. pp5-17

[Hankamer 1989] Hankamer, J (1989) Morphological Parsing and The Lexicon. In Marslen-Wilson, W [ed]*Lexical Representation and Process.* The MIT Press, Cambridge, Mass. pp392-408.

[Kjellmer 1991] Kjellmer, G (1991) A mint of phrases. In G Aijmer and B Altenberg [eds] 1991 *English Corpus Linguistics. Studies in Honour of Jan Svartvik.* Longman, London. pp111-127.

[Kurtböke 1998a] Kurtböke, P (1998a) Delexicalized verbs in Turkish from a corpus perspective. Proceedings of the 9th International Conference on Turkish Linguistics. University of Oxford, 12-14 August 1998.

[Kurtböke 1998b] Kurtböke, P (1998b) A Corpus-Driven study of Turkish-English language contact in Australia. Ph.D Thesis. Monash University, Melbourne.

[Kurtböke/Beadle *in progress*] Kurtböke, P and J Beadle (*in progress)* Turkish-English Bilingual Dictionary.

[Nattinger/DeCarrico 1992] Nattinger, J R and J S DeCarrico (1992) *Lexical Phrases and Language Teaching.* Oxford University Press,

[Pawley/Snyder 1983] Pawley, A and F Snyder (1983) 'Two Puzzles for Linguistic Theory: native-like selection and nativelike fluency'. In J C Richards and R W Schmidt [eds] *Language and Communication.* Longman, London. pp191-226.

[Partington 1993] Partington, A (1993) Corpus Evidence of Language Change – The Case of the Intensifier. In M Baker *et al* (eds) *Text and Technology. In Honour of John Sinclair.* John Benjamins, Amsterdam. Pp177-192.

[Pierce 1961] Pierce, J (1961) A Frequency count of Turkish Affixes. *Anthropological Linguistics* 3:9 31-42.

[Pierce 1962] Pierce, J (1962) Frequencies of occurrence for affixes in written Turkish. *Anthropological Linguistics* 4:6 pp30-41.

[Quirk 1992] Quirk, R 1992 On Corpus principles and design. In J Svartvik (ed) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August 1991. Mouton de Gruyter, Berlin. Pp457-469.

[Sinclair 1991] Sinclair (1991) *Corpus Concordance Collocation* Oxford University Press, Oxford.

[Sinclair/Renouf 1988] Sinclair, J and A Renouf (1988) A Lexical syllabus for language learning. In R Carter and M McCarthy (eds) *Vocabulary and Language Teaching.* Longman, London. pp140-160.

[Storrer/Schwall 1995] Storrer, A and U Schwall (1995) Description and Acquisition of Multiword Lexemes. In Steffens, P *Machine Translation and the Lexicon.* Springer, Berlin. pp35-50.

[Van Der Meer 1998] Van der Meer, G (1998) Collocations as one particular type of conventional word combinations. Their definition and character. In T Fontenelle *et al*. [eds] 1998 *EURALEX '98 Proceedings.* University of Liège, Belgium. pp313-322.