Morphological preprocessing or parsing: where does semantics meet computation?

Cem Bozşahin
Middle East Technical University (ODTÜ), Ankara, Turkey
bozsahin@metu.edu.tr

ABSTRACT

The paper argues for two points in relation to Turkish NLP: (i) we are better off developing and using research methodologies and tools that are not language-specific, although the models built with these methods and tools must certainly exploit language-specific thinking or technology. One way to do is to collect distributional data at the level of morphemes. (ii) we need to incorporate semantics into the picture somehow, otherwise what we do is form recognition, or contextually deprived (or dissituated) form production. The last point raises problems from the world's morphologies (and from Turkish morphology in particular) for the current state of art in NLP, where morphological processing is usually separated from syntactic processing for practical reasons. There is no semantic motivation to separate morphological processing from syntactic processing. In fact, semantic aspects indicate that we should integrate them. I will mention some attempts at the problem and suggest some lines of research.

Keywords

morphology, parsing, semantics, syntax

1. INTRODUCTION

We have enough evidence to think of morphological processing as part of syntax. Although it is common to separate the issues in the current state of art to simplify processing, e.g. word-level processing first, then syntactic processing, this practice is based on the unsubstantiated belief that adding morphemes to syntactic processing would be computationally more demanding. Showing the consequences of not adding morphemes to the syntactic process is the main purpose of this short paper. We shall also see that morphemic processing at the syntax level is poor mainly because we have been collecting statistics at the word level rather than morpheme level, although abundant evidence for morpheme semantics is there in the data once we look for it.

Even for morphologically less complex languages, this approach proves very useful; see [3] for an early attempt in English. I will concentrate on Turkish in the current paper, with some examples from other morphologically complex languages to promote the stance in (i–ii) of the abstract.

2. WORDS AND PREPROCESSING

As cases of the two points in the abstract, I exemplify from two languages. I will subsequently explain how these aspects bear on morphological processing in Turkish.

Anderson describes in [1] K^{w}ak$^{w'}$ala's morphology, where some inflections of a noun are attached to the preceding word: "[..] the inflectional markers for case, deictic status, and possesor of every NP are found not within that NP itself but rather on the preceding element of the sentence—whatever that may be, regardless of its grammatical relation to the NP in question:"(p.19)

Below is an example from Anderson. For example, the subject/article marker for 'expert' is on the verb for guides, and the morphological case of the word for harpoon (instrumental) is on the word for hunter.

(1) nanaqəsil-ida iʔgəl'wat-i əlewinux̣^{w}a-s-is
guides-SBJ/ART expert-DEM hunter-INSTR-his
mestuwi la-x̣a mig^{w}at-i
harpoon PRE-OBJ/ART seal-DEM
'An expert hunter guides the seal with his harpoon.'
[1]; p.18

As Anderson shows, the space-separated words in the example are morphological and phonological words. The state of affairs seems to cause no computational difficulty to speakers of K^{w}ak$^{w'}$ala because there is always an element preceding a noun in K^{w}ak$^{w'}$ala; it is a verb-subject-object language. That element need not be a noun, which further complicates the picture for doing word-based morphological analysis before we start syntactic analysis (parsing). For example, we would not know until the syntactic stage whether to follow nominal morphology on a verb, noun or other lexical classes, or to mark a word as presyntactically verbal or nominal, unless we repeat all nominal morphology on all lexical classes. We can of course do that at the expense of having a baroque description of K^{w}ak$^{w'}$ala, but now the spirit of morphological preprocessing, which is to aid parsing by reducing the number of alternatives for parts of speech, is no more useful as a preparsing device than it would be as part of some n-gram trained model of part of speech selection at the syntactic stage.

The English plural suffix forces us to consider alternatives to preprocessing as well. When the plural-marked noun is

within scope of intersective adjectives such as *green* and *four*, the plural takes word-internal scope: *green* [*truck*]*s*. With nonintersective adjectives, the plural has phrasal scope: [*toy truck*]*s*, [*alleged thiev*]*es* (see [4] for a detailed description and a word-based solution, and [2] for an alternative solution). This is crucial for semantics because *toy trucks* is not an intersection of set of toys and nonsingleton sets of trucks (that would be the morphologically-scoped reading of the plural), it is the nonsingleton set of toy trucks.

We are, then, in a situation where we could do preprocessing in English if we wished, but that preprocessing can be rendered redundant if we leave the disambiguation of plural's scope to parsing (by collecting statistics about English morphemes rather than word forms), otherwise we will be overextending our morphological preprocessor's work to handle word bigrams, trigrams and more, because we also have *alleged night thieves*, *alleged tanker oil thieves* etc. (these examples are real; they are found in a google search).

The situation is even more severe in morphologically richer languages. Turkish possessive markers are four-way ambiguous, and the only way to figure out the right scope is to look at syntactic structure. Examples below are from [8]. The structures assumed by Göksel are in primes. The coindexation possibilities show that they must be distinguished structurally.

(2) a. sev-di-k-ler-imiz
like-T-REL-PLU-POSS.1p
'those who we like/liked'

a′ [SUB$_j$ —$_i$ (OBJ) sev-di-k-ler$_i$-imiz$_j$]

b. sev-en-ler-imiz
like-REL-PLU-POSS.1p
'those who like/liked us'

b′ [—$_i$ (SUB) OBJ$_j$ sev-en-ler$_i$-imiz$_j$]

c. köpek sev-en-ler-imiz
dog like-T-REL-PLU-POSS.1p
'those among us who like/liked dogs'

c′ [—$_i$ (SUB) OBJ sev-en-ler$_i$-imiz$_i$]

Another point of preprocessing is to prepare what is relevant to syntax and parsing. One way to do this is to group derivational morphemes and inflectional ones as for example [15, 10, 7] do (there are exceptions to grouping; see e.g. [16]). We see alternating groups of suffixes in causatives and other suffixes: *yap-tır-ım* make-CAUS-DER 'sanction' versus *kovala-t* dismiss-DER-CAUS 'cause to be chased' and *uyu-klama-yış* sleep-DER-NEG-DER 'alertness', 'vigilance' or 'not sleeping'. Once again, these distributional regularities are available to a syntactic parser once we collect the statistics over the morphemes, rather than some preconceived groupings in word forms.

3. TURKISH MORPHOLOGY MEETS SEMANTICS

Structural ambiguities do not only arise from inflections. In the following examples, we see both kinds. Thus it is important that we keep the derivational morphemes in mind when we collect statistics.[1]

[1]If [6] are right, the inflection-derivation distinction is arbitrary anyway.

(3) a. [Çok uzun kol]-lu gömlek
Very long sleeve-COM shirt
'Very-long-sleeved shirt'

b. Atlet-ler-in [en iyi]-si
Athlete-PLU-GEN.3s most good-POSS.3s
'The best of the athletes'

The infamous *-ki* suffix presents another problem in Turkish. Its one aspect, that it can repeat itself indefinitely, is handled nicely by the finite-state preprocessors. (Those who do morphology by morpholexical rules are forced to take a finite closure of the problem; see [17] for some discussion.)

Its semantics shows that it is best left to a parser. It can be attached to case-marked nouns whose case relation is one of possession, time, or place (i.e., the genitive and the locative), e.g. *ev-in-ki* (house-GEN-ki 'the one of the house') and *ev-de-ki$_i$ çocuk$_i$* (house-LOC-ki child 'the child in the house'). We thus get relative pronoun and relative adjective interpretations. *Ki*'s morphological effect is to create a nominal stem on which all inflections can start again. As [11] noted, there is no upper bound on this process of relativization (e.g. *ev-i-nde-ki-ler-in-ki-ler-de-ki*).

The semantics of both structures are clear: the relative adjective is coreferential with the next NP (4a), and the relative pronoun cannot be coreferential (4b).

(4) a. Ev-de-ki$_{i/*j/*k}$ küçük çocuk$_i$ fare-yi$_j$ gör-dü.
house-LOC-ki little child mouse-ACC see-PAST
'The little child in the house saw the mouse.'
lit. 'The little child, the one in the house, saw the mouse.'

b. Ev-in-ki$_{*i}$ fare$_i$ gör-müş.
house-GEN-ki mouse see-PERF
'The one in/of the house saw a mouse.'
meaning, e.g. The cat of the house saw a mouse.

The semantic challenges posed by *-ki* to syntax-morphology communication are twofold: only the relative pronoun interpretation is available if other inflections follow *-ki* (5a–b), and semantics of earlier *-ki*'s are blocked by the last *-ki* (5c–d). Considering a reasonable assumption that these recursive structures must be derived syntactically (otherwise Turkish lexicon would have to be infinite), we would need some auxiliary mechanisms to derive only the possible readings from morphological bracketings. The semantic issues, then, appear to be independent of whether we take *-ki* as an affix [9] or a clitic [14].

(5) a. Ev-de-ki$_{*i}$-ler-de çocuk$_i$.
house-LOC-ki-PLU-LOC child
'The child is with the ones in the house.' Turkish
* for 'the children, the ones at the house'

b. Ev-de-ki$_{*i}$-ler-de para$_i$ az.
house-GEN-ki-PLU-LOC money few
lit. 'There is little money on the ones in the house.'
meaning, e.g. there is little money on the students in the house.
* for 'the money, the ones at the houses, is few'

c. Ev-de-ki$_{*i}$-ler-in-ki hasta$_i$.
house-LOC-ki-PLU-LOC-ki ill
lit. 'The one of the ones in the house is ill.'
meaning, e.g. the guest of the family in the house is ill.

d. Ev-de-ki$_{*i}$-ler-de-ki$_i$ hasta$_i$
house-LOC-ki-PLU-LOC-ki patient
lit. 'The patient, the one at the ones in the house'
meaning, e.g. the patient, the one at the flat of the family in the house
* for 'the patients, the ones in the house'

4. MORPHEMES AND PARSING

One aspect that makes preprocessing an attractive alternative is efficiency. Rather than dealing with an abundant number of morphemes in a string, grouping them before parsing seems tempting. However, if we want our parsers to deliver semantics as well as recognize structure of strings, we might be making the task harder for the parser by preprocessing, as we have seen. The concern for efficiency is of course very legitimate, and ultimately it is the decisive factor in technology (and also for any computationalist theory, which is promoted here). As a first attempt at keeping the statistics at the morpheme level shows [5], we may be able to parse wide-coverage with morphemes *and* deliver semantics quite efficiently. The system delivers predicate-argument structures of parsed strings.

One trouble with working with radically lexicalized predicate argument structures (as semantic representations) at the level of morphemes is that these structures are difficult to annotate therefore train on large scale. One way to ease this problem is to automatically derive the predicate-argument structures from dependency structures, which are much easier to annotate because they are intuitive.[2] We have the training for dependency structures at the word level [15, 7], and [12, 5] have tried to induce predicate-argument structures from dependency structures. This is where a morphological processor might help, not as a *pre*-processor to syntax but as a device for gathering data from wide variety of resources at the morpheme level; see for example [13, 18].

The semantics of morphemes can be statistically approximated from document use as well, as [18] does. Latent Semantic Analysis puts a web twist on the Wittgensteinian idea of taking the meanings as products of practice. This way of thinking will benefit from morphological processors as stand-alone systems. For parser training, the current paper argues that we must connect the output of such systems to lexicons and parsing models, rather than rely on preprocessing.

5. CONCLUSION

We may ask ourselves why we should change the practice of morphological preprocessing just because some rarely used word-internal recursion in Turkish and some marginal cases such as K$^{\text{w}}$ak$^{\text{w'}}$ala suggest that it may not be helpful. Besides the usual considerations of science versus technology, the fact remains that information which preprocessing provides to subsequent stages is available to a single-stage process if we compile our statistics on morphemes rather than word forms. All languages can benefit from this way of thinking, where parsing works more or less the same way, and morphology feeds the syntactic process not as a separate stage but in the form of statistical information available to parsing.

[2]The intuitions seem to be not reflected fully when dependencies are labeled. See [5] for some discussion.

I emphasize that I am not arguing against morphological processing as a useful device, but about what parsing really needs for training. It seems to be the combinatory knowledge of morphemes, not just for morphologically complex languages but for all languages. A word-based preprocessor would inevitably try to overextend itself to do that task. This burden is put on its shoulders by semantics, most notably by scope and by structural relations such as constituency and dependency.

6. REFERENCES

[1] S. R. Anderson. *A-Morphous Morphology*. Cambridge Univ. Press, 1992.

[2] C. Bozsahin. The combinatory morphemic lexicon. *Computational Linguistics*, 28(2):145–176, 2002.

[3] M. R. Brent. From grammar to lexicon: unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262, 1993.

[4] B. Carpenter. *Type-Logical Semantics*. MIT Press, Cambridge, MA, 1997.

[5] R. Çakıcı. *Wide-coverage Parsing for Turkish*. PhD thesis, University of Edinburgh, 2008.

[6] A. M. Di Sciullo and E. Williams. *On the Definition of Word*. MIT Press, Cambridge, MA, 1987.

[7] G. Eryiğit, J. Nivre, and K. Oflazer. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389, 2008.

[8] A. Göksel. Pronominal participles in Turkish and lexical integrity. *Lingue e Linguaggio*, 5(1):105–125, 2006.

[9] A. Göksel and C. Kerslake. *Turkish: A Comprehensive Grammar*. Routledge, London, 2005.

[10] Z. Güngördü and K. Oflazer. Parsing Turkish using the Lexical-Functional Grammar formalism. *Machine Translation*, 10:293–319, 1995.

[11] J. Hankamer. Morphological parsing and the lexicon. In W. Marslen-Wilson, editor, *Lexical Representation and Process*. MIT Press, Cambridge, MA, 1989.

[12] J. Hockenmaier and M. Steedman. CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):356–396, 2007.

[13] B. İlgen and B. Karaoğlan. Investigation of Zipf's 'law-of-meaning' on Turkish corpora. In *22nd international symposium on Computer and information sciences (ISCIS)*, pages 1–6, 2007.

[14] J. Kornfilt. *Turkish*. Routledge, London, 1997.

[15] K. Oflazer. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544, 2003.

[16] H. Sak, T. Güngör, and M. Saraçlar. Resources for Turkish morphological processing. *Language Resources and Evaluation*, 45(2):249–261, 2011.

[17] O. Sehitoglu and C. Bozsahin. Lexical rules and lexical organization: Productivity in the lexicon. In E. Viegas, editor, *Breadth and Depth of Semantic Lexicons*. Kluwer, 1999.

[18] V. T. Turunen and M. Kurimo. Indexing confusion networks for morph-based spoken document retrieval. In *Proc. of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pages 631–638, 2007.