AN ADAPTABLE MORPHOLOGICAL PARSER FOR AGGLUTINATIVE LANGUAGES

Lomonosov Moscow State University

Marina Ermolaeva

Department of Theoretical and Applied Linguistics
Faculty of Philology
✉ marinkaermolaeva@gmail.com

Abstract

- The poster reports the ongoing work on creating a multi-language parser, suitable for languages with agglutinative morphology.
- A hybrid approach involving methods typically used for non-agglutinative languages is proposed.
- We explain the design of a working prototype for inflectional nominal morphology and demonstrate its work with implementations for Turkish language (Altaic, Turkic) and Buryat language (Altaic, Mongolic).

1. Introduction

1.1. The main idea

- A simple way to perform morphological parsing: list all possible forms of each word. This method yields good results for non-agglutinative languages (Segalovich 2003).
- Finite-state machines (FSMs) can take care of an infinite set of possible word forms. They are widely used for agglutinative languages, including Turkish (Eryiğit&Adalı 2004, Şahin et al. 2013 etc.)
- In order to achieve relative language independence, the proposed approach combines both methods.

1.2. Processing direction

- Unlike most systems, starting with (Oflazer 1994), we apply the right-to-left parsing method (cf. (Eryiğit&Adalı 2004)) to simplify processing words with unknown stems.

1.3. Progress so far

- The proposed system is multi-language (cf. (Akın&Akın 2007; Arkhangelskiy 2012)).
- The working prototype is currently restricted to nominal inflectional morphology.
- The system does not disambiguate yet; in case of ambiguity, the output includes all plausible parses.

2. Turkish challenges

- The complexity of Turkish morphology is easily perceptible in nouns.
- Hyphenless compounding is productive.
- Due to the vowel harmony and assimilation rules, most suffixes have multiple allomorphs distributed according to their phonological context.
- A nominal stem receives number, possession and case suffixes.
- Nominal forms can be modified further, yielding predicates or adverbial forms.

(1) ev-ler-imiz-de-ymiş-ler[1]
home-PL-P1PL-LOC-COP.EV-3PL
Apparently they are/were at our homes.

- The suffix *-ki* can be recursively attached to a nominal form with a genitive or locative marker[2]:

(2) ev-de-ki-ler-in-ki
home-LOC-KI1-PL-GEN-KI2
the one belonging to those at home

Footnotes

1. Examples (1) and (2) are from (Göksel and Kerslake 2005).
2. According to Hankamer (2004), *-ki* has different properties when attached to a locative form and to a genitive form; therefore, two separate *-ki*'s are postulated. In this paper, they are referred to as KI1 and KI2 respectively.
3. The Turkish implementation employs a lexicon of 16000 nominal and adjectival stems, originally from http://www.fen.bilkent.edu.tr/~aykutlu/sozluk.txt.
4. Buryat examples are presented both in the traditional Cyrillic orthography and in Latin transcription; the actual implementation works with the former.

3. System design

3.1. Data representation

- Long morpheme sequences are split up. Grammatical categories are arranged into a set of slots, each containing categories with strictly fixed order(s):
  - two stem slots
  - noun inflection
  - noun loop (the recursive *-ki*)
  - nominal verb suffixes (e.g. copulas)
- The number and order of categories within slots can be changed without modifying the system itself.
- For each slot, all possible suffix sequences are obtained. Suffix compatibility inside a slot is checked at this stage.
- Suffix sequences and stems are inverted and stored in letter trees. Allomorphs are treated as separate entries.

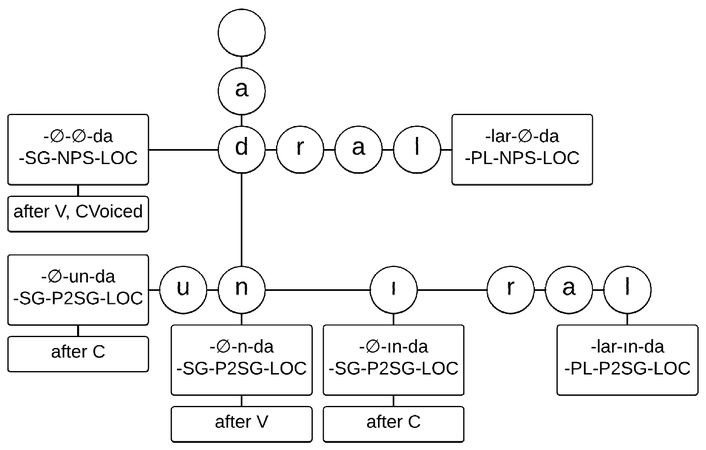

Figure 1. A fragment of the noun inflection tree

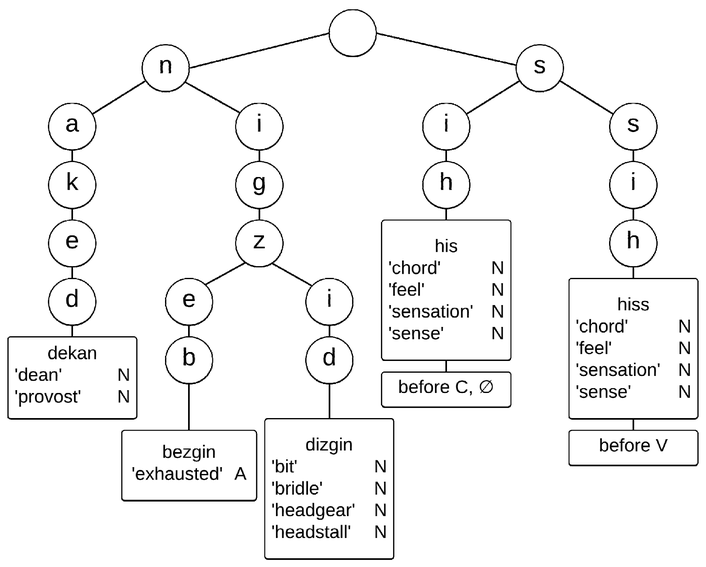

Figure 2. A fragment of the lexicon tree

3.2. Parsing process

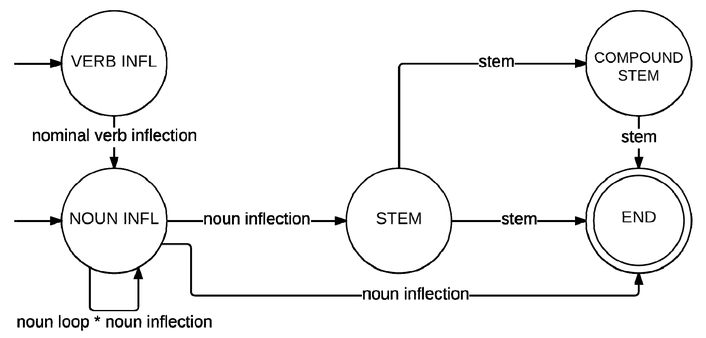

Figure 3. The FSM used for transitions between slots

- Each transition corresponds to a sequence of suffixes rather than to a single suffix.
- Linguistic information is only used between slots.

3.3. Examples

(3)
input: *adamdı*
decision: single stem
output:
1. adam-∅-∅-∅-dı-∅
   man-SG-NPS-NOM-COP.PST-3
2. ada-∅-m-∅-dı-∅
   island-SG-P1SG-NOM-COP.PST-3

(4)
input: *fefe*
decision: unknown stem
output:
1. fef-∅-∅-e
   FEF-SG-NPS-DAT
2. fef-∅-∅-e-∅-∅
   FEF-SG-NPS-DAT-COP.PRS-3
3. fefe-∅-∅-∅
   FEF-SG-NPS-NOM
4. fefe-∅-∅-∅-∅-∅
   FEF-SG-NPS-NOM-COP.PRS-3

4. Evaluation

- The evaluation method (Paroubek 2007) takes ambiguity into account.
- First, precision (P) and recall (R) values for each word w_i in the test sample are obtained:

$$P(w_i) = \frac{t_i \cap r_i}{t_i} \qquad R(w_i) = \frac{t_i \cap r_i}{r_i}$$

where t_i is the number of parses for w_i output and r_i is the number of correct parses.

- After that, mean values for the whole sample are calculated.
- We accept a parse if the stem-suffix boundary is determined correctly and all nominal inflectional suffixes are properly labelled.
- Results: precision ≈94%; recall ≈96%.

5. Buryat challenges

- What if we choose a non-Turkic language?
- Like that of Turkish, Buryat morphology is agglutinative and suffixal.
- Buryat poses more challenges in some respects.

(5) a. таабар-∅-иин-∅[4]
ta:bər-∅-in-∅
riddle-SG-GEN-NPS

b. таабари-∅-∅-мни
ta:bəri-∅-∅-mni
riddle-SG-NOM-P1SG

(6) a. гэр-∅-эй-∅
gэr-∅-e-∅
house-SG-GEN-NPS

b. гэр-∅-∅-ни
gэr-∅-∅-ni
house-SG-NOM-P1SG

- A small part of Buryat morphology has been modelled. No language specific modifications were done to the parser itself.
- Evaluation: precision ≈91%; recall ≈96%.

6. Future work

- The natural to-do list: other parts of speech, derivational suffixes, disambiguation.
- Implementing new languages may require a more flexible slot system. This can be achieved by designing a near-universal slot system or by deriving it automatically from a corpus.
- DIRETRA, an engine for Turkish-to-English word-for-word translation reflecting morphology, is being developed on the base of the parser (Aksënova&Ermolaeva, in prep.)

input	adamlarınkiler
parser output	man-PL-GEN-KI2-PL
DIRETRA output	ones.owned.by.men

Table 1. A sample DIRETRA output

Abbreviations

1 – first person, 2 – second person, 3 – third person, COP.EV – evidential copula, COP.PRS – present tense copula, COP.PST – past tense copula, DAT – dative, GEN – genitive, KI1 – locative *-ki*, KI2 – genitive *-ki*, LOC – locative, NOM – nominative, NPS – non-possession, P – possession, PL – plural, SG – singular.

References

Ahmet Afşın Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source NLP framework for Turkic Languages.

Alëna Aksënova and Marina Ermolaeva. In prep. DIRETRA, a customizable direct translation system: first sketches. In: Proceedings of Translata II.

Timofey Arkhangelskiy. 2012. Printsipy postrojenija morfologicheskogo parsera dlja raznostrukturnyx jazykov [Principles of building a morphological parser for different-structure languages]. Abstract of thesis cand. phil. sci. Moscow.

Gülşen Eryiğit and Eşref Adalı. 2004. An Affix Stripping Morphological Analyzer for Turkish. In: IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, 299-304.

Aslı Göksel and Celia Kerslake. 2005. Turkish: A Comprehensive Grammar.

Jorge Hankamer. 2004. Why there are two ki's in Turkish. In: Imer and Dogan, eds., Current Research in Turkish Linguistics, Eastern Mediterranean University Press, 13-25.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. In Literary and Linguistic Computing, vol. 9, no. 2, 137-148.

Patrick Paroubek. 2007. Chapter 4 - Evaluating Part Of Speech Tagging and Parsing. In: Evaluation of Text and Speech Systems, eds. Laila Dybkjær, Holmer Hemsen, Wolfgang Minker, series: Text, Speech and Language Technology, vol. 36, Kluwer Academic Publisher, 97-116.

Muhammet Şahin, Umut Sulubacak and Gülsen Eryigit. 2013. Redefinition Of Turkish Morphology Using Flag Diacritics. In: Proceedings of the Tenth Symposium on Natural Language Processing (SNLP-2013).

Ilya Segalovich. 2003. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. MLMTA, 273-280. CSREA Press.