Case Inflection in Koalib: Discovering the Rules

Georgi Boychev
Independent Researcher
georgy_boytchev@abv.bg

Claire Gardent
CNRS/LORIA, Nancy (France)
claire.gardent@loria.fr

Nicolas Quint
LLACAN-UMR8135(CNRS/INALCO/PRES-Paris-Cité)
quint@vjf.cnrs.fr

The Koalib language belongs to the Heiban group, Kordofanian branch, Niger-Congo family (Schadeberg, 1981) and is spoken by approximately 100 000 persons in the Nuba Mountains, Sudan (Quint, 2010). It has two grammatical cases – subject case (S) and object case (O) (Quint, unpublished). Based on a corpus of 1 200 Koalib nouns, Quint found that 75% of the nouns are marked for case by either:

- adding a suffix (35%),
  e.g. **kàllány** (S) 'axe' > **kàllányé** (O),
- changing the tonal pattern (6%),
  e.g. **kwìcì** (S) 'person' > **kwícì** (O),
- or both (34%),
  e.g. **káaŋàl** (S) 'sheep' > **kàaŋálè** (O).

Quint (unpublished) discovered five case-inflection rules which cover around 35% of the data with 82% accuracy. Finding the rest of the rules appeared to be too complex for manual analysis.

This study aims to discover a larger set of rules with machine learning methods. We trained a RIPPER classifier (Cohen, 1995) on the corpus used by Quint (unpublished) and extracted the rules from it.

A classifier is an algorithm which predicts the classes of items according to some training data. The training data contains a list of items and their correct classes. Each item is described by a set of features. The classifier first builds a model from the training data. The model associates each class with its most probable features. Then, it uses the model to predict the classes of new items, given their features.

In our case, each item is a Koalib noun and its features are the phonemic, syllabic and tonal properties of its subject form. The classes describe the exact change from subject form to object form with a special class for cases with no change. For instance, the class "ADD-H + a" denotes the class of nouns whose object form is obtained by adding the suffix ***-a*** with a high tone to the subject form, e.g. **kòrkókkè** (S) 'tortoise shell' > **kòrkókkèá** (O).

The trained model is a list of rules which, if applied in order, predict the class of any Koalib noun. Each rule contains a set of features on its left-hand side and the most probable class on its right-hand side:

- (LAST-TWO-TONES=LH) and (LAST-PHONEME=V) -> CLASS=ADD-H + ŋe,
  e.g. **kòţţó** (S) 'gourd' > **kòţţóŋé** (O).

We evaluated the rules by measuring the prediction accuracy (the number of correct predictions divided by the total number of predictions) of the model on unseen data. Our best model achieved 66.25% accuracy on the whole dataset. However, some of its rules have poor accuracy and cover only few nouns. These can be filtered out to obtain more reliable rules, while slightly reducing their coverage. It should be noted that the size of the corpus may have been too small to discover more subtle patterns, assuming that such even exist. Thus, it might be worthwhile to examine if more rules can be discovered with a larger corpus as future work.