Designing finite-state morphological transducers for Kypchak languages

This paper describes the development of free/open-source finite-state morphological transducers for several Northwestern Turkic (or Kypchak) languages. The transducers function to provide morphological analyses of surface forms as well as produce surface forms from morphological information, and can be repurposed as spell-checkers for these under-resourced languages. The finite-state toolkit used for the work is the Helsinki Finite-State Toolkit (HFST) (Linden et al., 2011).

Work is ongoing on morphological transducers for seven Kypchak languages: three Southern Kypchak languages (Kazakh, Karakalpak, and Nogay), two Northern Kypchak languages (Tatar and Bashqort), one Western Kypchak language (Kumyk), and Kyrgyz. This paper describes how the development of a transducer for each subsequent language took less time than the previous one.

An evaluation is presented for the transducer of one language from each of the four branches of Kypchak (Kazakh, Tatar, Kumyk, and Kyrgyz). Each transducer has production-level coverage—around 90%—on freely available corpora of the languages, and high precision (the number of the analyses given for a form that are correct) and recall (the percentage of possible correct analyses for a form that are provided by the transducer) over a manually verified test set.

The computational implementation of the morphotactics and morphophonology of the languages is described. Emphasis is placed on the types of differences that exist between the morphologies of these languages, as well as on previously undocumented morphological patterns in these languages that were discovered through the development of the transducers.

References

Linden, Krister, Silfverberg, Miikka, Axelson, Erik, Hardwick, Sam, & Pirinen, Tommi (2011). *HFST—Framework for Compiling and Applying Morphologies*, vol. Vol. 100 of *Communications in Computer and Information Science*, pp. 67–85. ISBN 978-3-642-23137-7.