Normalization of dialects/variants using FST technology

Normalization of dialects and variants is an interesting are of the NLP. It is used for Information Retrieval on historical texts dialectal texts, preprocessing of dialectal variants (before processing these corpora using standard NLP tools), conversion speech-text...

There are two main approaches: rule-based and data-driven. Finite-state technology is adequate for both approaches.

In the first approach a grammar is written based on a linguistic description of the changes among the variant and the standard (or pivot) language. Toolkits for describing phonological/morphological changes are used. Foma will be the toolkit we will use for this.

The data-drive approach is based in a parallel list of words (pairs the equivalent words in the variant and the standard language). The toolkit is oriented to learn from this list and to generalize the changes. The noisy-channel model is very popular for this kind of task. Phonetisaurus will be the toolkit we will use for this.

SLIDES
Basic material: http://foma.sourceforge.net/lrec2010/index.html
But adapted to less-resourced languages and normalization

TOOLKITS:
Rule-based (foma): (http://code.google.com/p/foma/)
Data-driven approach: Phonetisaurus (http://code.google.com/p/phonetisaurus/)

PROJECT: Development of a simple normalization-tool for a language/dialect
(it would be interesting if the students propose a real problem where a list or a formal description is available)