Cheap resource development

for a resource-light morphological tagger/analyzer

Anna Feldman and Jirka Hana

Constraints

- Avoiding linguistic expertise in the target language
- Limiting:
  - general linguistic expertise
  - (native) expertise in target language
  - training time
  - total time needed

Human resources

Tagset, paradigms, word lists, ... created by

- a non-expert (not a linguist and not a native speaker).
- preferably a person
  - with knowledge of a related language or a language with similar inflectional categories e.g. for Russian, knowledge of Polish is ideal, but even knowledge of German is better than just knowledge of English (it has gender and case)
  - who has created resources for another language before

Tagset

- The problem is usually ignored
- Surprisingly hard to create
- Providing template saves time on many arbitrary decisions (e.g. is numeral a POS?, how to denote it, etc)

Other things

- Exploring alternative/simpler ways in which to specify paradigms (usually the grammar books are "messy")
- Measuring impact of time/expertise directed in different directions
- Experimenting with Romanian, Lithuanian, Belorussian, Old Czech
- Course at ESSLLI 2010

A totally different thing – Annotating learners corpora