Getting more data – Schoolkids as annotators
Jirka Hana & Barbora Hladká
Charles University in Prague

Introduction

We present a new way to get more morphologically and syntactically annotated data.
Keywords: annotated corpora, morphology, syntax, crowdsourcing.

An idea – A case study

1. Select the annotation framework for the target corpus.
   > the Prague Dependency Treebank
2. Design and implement an annotation editor tailored to school children to involve them in text annotation. > Capek
4. Transform their analyses into the chosen annotation framework.

Sentence diagramming at Czech schools

▶ Children at the elementary and high-school level are required to parse sentences into dependency trees.
▶ Is sentence diagramming taught in your country?

Computational linguistics goes to Czech schools

U nedalekého potoka už rozkvetly žluté blatouchy
By near stream already blossomed-out yellow kingcups
‘By a stream nearby, yellow kingcups have blossomed out.’

The STYX system

▶ 11K sentences with their morphological and syntactic analysis.
▶ Charon to select sentences from the database.
▶ Styx to analyze sentences and to check the analysis.

Czech school analyses and their processing

Subject-predicate transformation Splitting multi-token nodes with preposition

Preliminary results

CoNLL 2007 Unlabelled attachment score (UAS) measures the percentage of tokens with the correct/same head (i.e., it evaluates the structural correctness disregarding syntactic labels) as the agreement metric:

- UAS(parser vs. gold-standard) = 89%
- UAS(teacher1 vs. teacher2) = 92%
- UAS(children1 vs. children2) = 85%
- UAS(transformation of teacher1’s annotation = 81% vs. gold-standard)