ADEPT

Charlotte Gooskens, CLCG

CLARIN-NL Workshop

19 Feb 2010

ADEPT

- Assaying Differences via Edit-Distance of Pronunciation Transcriptions
- Vision: enable non-programmers to measure pronunciation dialect differences
- 6 month, € 60K development costs
- Based on L04, open source C programs
  - In active use since about 2004
  - About 20 users world-wide

Team

- Charlotte Gooskens, RuG, co-ordinator
  - User of L04 for dialectology, language contact study
- Peter Kleiweg, RuG, developer (also developer of L04)
- Jan Pieter Kunst, Meertens, CLARIN center expert
- John Nerbonne & Martijn Wieling, RuG, advising & designing

Edit Distance of Pronuciation Transcription

æ		f	t	ə	r	n	u	n
æ	ə	f	t	ə		n	✶	n
	1				1		1	

- At aggregate (site) level this rough measure is adequate
- Techniques available to determine segment distances empirically
  - Finer measures interesting in determining likely sound correspondences

Stumbling Blocks for Users

- Installation & invocation – UNIX facilities that can be linked via scripts found forbidding
- Checking validity of IPA transcriptions
  - Many data sets use “convenience” encodings
  - ‘@’ or ‘∂’ for ə

- We aim to solve the first problem via construction of a web application
- We aim to solve the second via a tool for checking the validity of IPA (perhaps later)

Technically

- Input requires IPA, e.g. in UTF-8, UTF-16
  - X-SAMPA also OK
  - All encodings weak in transparency
    - “obvious” choices (from keyword) often incorrect, e.g. keyboard ‘N’ (U +4E) for uvular nasal instead of unicode (U+0274)
    - Unicode decomposition =/= IPA decomposition
      - [ç] =/= IPA palatal stop + cedilla (not a phonetic diacritic)
  - We’ll *report* on problems
- ISO-cat has little to say about phonological and phonetic categories, extension required

Further Analysis

- Planned work includes providing a consistency measure, Cronbach's α
- L04 contains links to software for further analysis and visualization, incl. MDS, (bootstrap) clustering (of various sorts), several sorts of maps
  - Probably outside the CLARIN-NL standardization charter, but we'll try to include some unofficially

Work Plan

- Need: letter of approval,
- Begin asap after arrival of letter
- + 2 mon.:
  - ISOcat additions, mapping, doc on IPA in Unicode
  - Design of web application (Pylons)
- + 3 mon.: demo scenario, metadata
- + 4 mon.: prototype on server, testing
- + 6 mon. User course (tutorial), offer to LOT, Methods XIV (London); doc. on requirements & infrastructure