Metadatenkonvertierung mit EUCALYPTUS und COALA

von SpeechRecorder nach CMDI

Bernhard Jackl
jackl@phonetik.uni-muenchen.de

31. März 2014

Wie speichern wir Metadaten zum Corpus ab?

- Tabelle (auch von Menschen gut lesbar)
- (rel.) Datenbank (sehr mächtig!)
- XML (sehr schlecht menschenlesbar, gut maschinenlesbar)
- ...

Alte Corpora haben möglicherweise ganz andere Metadaten-Formate als die aktuellen, je nach Projekt!

CLARIN

- Infrastrukturmaßnahme
- soll Auffinden von Corpora und Tools erleichtern
- Frühere Metadaten-Initativen: Dublin Core ('98), IMDI (2000), ...
  - entweder sehr wenige, allgemeine Einträge (DC), oder sehr spezialisiertes Format (nur für Multimedia und multimodale Sprachresourcen bei IMDI)
- Neuer Standard ist CMDI (XML)

CMDI

- Einführung von Profilen
  - bestehen aus Components (die selbst aus weiteren C. und atomaren Elements bestehen)
  - registriert in der ComponentRegistry
  - Begriffe definiert in ISOcat
- wenn abwärtskompatibel können Profile erweitert werden
- Profile des BAS:
  - media-corpus-profile (v1.0)
  - media-session-profile (v1.2)
- Problem: Vielzahl an Profilen, Gefahr der Zersplitterung

Wie erstellen wir CMDI?

- per Hand (sehr mühsam!)
- mit Arbil (Java CMDI-Editor und -Viewer, je nach Größe des Corpus viel Arbeit)
- mit einem Script, das direkt CMDIs erzeugt (besser, muss aber für jedes Corpus neu erstellt werden)
- Zwischenformat erstellen und dieses zu CMDI konvertieren

Art des Zwischenformates

- 5 Tabellen (CSV Format, UTF-8)
  1. Actors (Teilnehmerin, Sprecherin, . . .)
  2. Sessions
  3. Bundles (einzelne Aufnahme, bei uns: ein Prompt)
  4. MediaFiles (Signaldateien, Teil eines Bundles)
  5. WrittenResources (Annotationen, Teil eines Bundles oder einer ganzen Session)
- Können erstellt werden:
  - per Hand (Excel, OpenOffice.org, . . .)
  - mit einem (Shell-) Script
  - als Export aus einer Datenbank

EUCALYPTUS

- ist spezialisiert auf SpeechRecorder
  - parst Projektdatei von SpeechRecorder
  - analysiert aufgenommene WAV-Dateien
- kann die erforderlichen 5 Tabellen erstellen
- erzeugt auch einen sog. Wrapper, der das Programm aufrufen kann, das die CMDIs generiert (COALA)

COALA

- umfangreiches Perl-Programm
- erzeugt aus den 5 Tabellen:
  - 1 unfertige Corpus CMDI (muss ausgefüllt werden)
  - mindestens 1 Session CMDI (für jede Session eine; bereit für den Ingest ins Repository)
- kann Session-CMDIs validieren
- ermöglicht Updates bestehender Daten (Einschränkung: Die Daten müssen im BAS Repository liegen)

COALA (Cont.)

- kann auch als Webservice genutzt werden (ähnlich wie WebMAUS)
- COALA und seine Abhängigkeiten müssen dann nicht installiert werden
- momentan in der Testphase (`curl -X GET http://clarin.phonetik.uni-muenchen.de/BASWebServices/services/help`)
- Web App ist in Planung

COALA übernimmt nicht die ganze Arbeit!

Nur eine Arbeitserleichterung!

Vielen Dank
für Eure Aufmerksamkeit!