Spoken language corpora

Martin Volk
Universität Zürich

Topics

- London-Lund-Corpus
- Spoken part of BNC
- Verbmobil Corpus
- Spoken language in newspaper corpora
- Other spoken corpora (in TIGERSearch)
  - SwitchBoard Corpus
  - Christine Corpus

Spoken language corpora

- are a central resource for speech recognition and speech generation systems
- need to cover all interesting phenomena of a language
- but also important in:
  - dialect research
  - language preservation efforts
  - psycholinguistics, logopedics, neurology, ...

Spoken language corpora

- DE: *Sprachdatenbank*
- comprise: audio-data and symbolic descriptions

Spoken language corpora

- Parameters:
  - Number of speakers (age, gender, social status, dialects)
  - Situation (monologue vs. dialogue, free or moderated)
  - Recording quality
  - Level of annotation

Verbmobil

- Large-scale German project for speech-to-speech machine translation (1993-2000)
- Corpus:
  - 1350 speakers in 1950 dialogues with 81'000 turns
  - Business Style conversation (role-playing) for appointment scheduling
  - different microphones and recording technologies
  - annotation: orthographic with a rich inventory for spontaneous speech (hesitations, corrections, noise)
- Demo: http://www.ims.uni-stuttgart.de/projekte/verbmobil/Dialogs/

Verbmobil (from Wahlster (ed.) 2000)

- The so-called partitur format used for the Verbmobil speech corpora orchestrates 15 strata of annotations:
  - two transliteration variants,
  - lexical orthography,
  - canonical pronunciation,
  - manual phono-logical segmentation,
  - automatic phonological segmentation,
  - word segmentation,
  - prosodic segmentation,
  - dialog acts, noises, superimposed speech, syntactic category, word category, syntactic function, and prosodic boundaries.

Verbmobil

- The multilingual Verbmobil corpus includes
  - bilingual dialogs (from Wizard-of-OZ experiments, face-to-face dialogs with human interpreters, or dialogs interpreted by various versions of Verbmobil) and
  - aligned bilingual transliterations.
  - Three treebanks for German, English and Japanese have been developed with annotations on three strata: morpho-syntax, phrase structure, and predicate-argument structure.

Transkription

- Kapitel Gesprächsanalyse der Studien-CD Linguistik
  - Umgang mit Gesprächen S. 3/4
  - Transkriptionssystem S. 4/4
  - Perspektiven der Gesprächsanalyse S. 2/3
  - Gesprächseröffnung (Dialekt) S. 3/6
  - Formen des Sprecherwechsels S. 4/4

Gesprochene Sprache in geschriebenen Texten

- Interviews
- Wörtliche Rede
- Reden (Politik, Lehre)
- verschriftlichte Radio-Reportage
- Comics

Gesprochene Sprache im CZ-Korpus

- Interviews: Suche 'INTERVIEW' oder 'CZ:'
  - Kurzvorstellung des Gesprächspartners
  - viele Fragen, Anreden
  - teilweise kurze Antwortsätze
  - Beispiele:
    - CZ: *Heißt das, daß Sie sich von der IBM-Welt abkehren?*
    - Wang: *Natürlich nicht. Beispielsweise ist OS/2 ...*
    - CZ: *Hat sich damit auch etwas an der Lizenzierungspraxis geändert?*
    - Wang: *Ja. Sie ist flexibler geworden ...*

Gesprochene Sprache im CZ-Korpus

- Unterbrechungen
  - Niedermaier: *Es gibt zur Zeit keine praktische und wirtschaftliche Alternative zu den relationalen Datenbanken ...*
  - CZ: *... Der Nutzen von ODBMS ist doch aber unbestritten ...*
  - Niedermaier: *... Ich bestreite den Nutzen ja auch nicht.*
- Metaphorik
  - CZ: *Wo **drückt** IBM **der Schuh** am meisten?*
  - *CZ: Kritiker sind der Meinung, Informix sei zwar auf der Client-Seite stark, bei den Servern jedoch **schwach auf der Brust**. Was sagen Sie dazu?*
  - *CZ: Compaqs Billigrechner der Prolinea-Serie haben einen Preiskampf im PC-Marken-Markt ausgelöst. Wie lange können Sie **das Gemetzel** selbst **durchhalten**?*

Gesprochene Sprache im CZ-Korpus

- Abschleifungen
  - CZ: *Herr Dreyer, während die Netzbetreiber Mannesmann und Telekom kräftig die Werbetrommel rühren, blieben die Diensteanbieter in der Öffentlichkeit bislang so gut wie unbekannt.* ***Woran liegt's?***
  - CZ: *Wie* ***steht's*** *derzeit mit der Qualität der vorhandenen digitalen Mobilfunk-Netze?*
  - CZ: *Wurde bisher verlangt, die Telekom soll dem Wettbewerb ausgesetzt werden, macht es jetzt den Eindruck als antworteten Sie: Wenn Ihr Wettbewerb wollt,* ***mach'*** *ich mit.*

Trigger für gesprochene Sprache

- Trigger-Wort (z.B. *Interview* oder *Eigenname* + *:*)
- Anführungszeichen (+ Komma/Fragezeichen + *sagte, teilte mit, erklärte*, ...)
- Apostroph + *s*

NZZ (April 1994): Literatur

- Ab und zu schafft er es, einzelne Wörter am Reis vorbeizuschieben, wir verstehen: "Wette, endlich Sonntag, Süden, zwei Stunden, Als Auto, schon gekauft."
- Als er fertig ist und den leeren Teller vor sich dreht, frage ich schnell, bevor er wieder loslegen kann: "Und was machst du, wenn du gewinnst?" Ein vorsichtiger Blick streift mich. "Geld kann man immer brauchen", sagt er kühl, "vielleicht verreise ich. Vielleicht wirklich weit weg." Schweigend steht er auf und kramt in seiner Hosentasche. "Viel Glück", wünsche ich. Da hellt sich seine Miene noch einmal auf. "Ja", lacht er, "Glück!" wirft ein paar zerknitterte Noten auf den Tisch und eilt weg.