AUVIS

Audiovisuelles Data-Mining am Beispiel der Ereignis-Segmentierung in multimodalen Sprachdaten

Einleitung

- Die vollständige manuelle Annotation von multimodalen Diskursdaten (Audio-/Videoaufnahmen von Kommunikationsereignissen) ist meist sehr aufwendig.
- Häufig stehen die dafür erforderlichen Ressourcen nicht zur Verfügung.
- Für viele Fragestellungen aus der Linguistik, Phonetik, Ethnologie, Psychologie und anderen Disziplinen ist eine solche vollständige Annotation nicht unbedingt erforderlich.
- Alternative zur vollständigen manuellen Annotation:
  - Schnelle Suche nach spezifischen Einheiten (Wort, Konstruktion, Sprechhandlung, Geste usw.) im nicht annotierten Teil eines Korpus mittels geeigneter Verfahren
  - Gezielte manuelle Analysen und Annotationen an den Fundstellen
  - Näherungsweise Generalisierungen über den nicht annotierten Teil des Korpus
- **Technische Ziele des AUVIS-Projekts**
  - Entwicklung von Verfahren zur automatischen Analyse von Audio- und Video-Daten (teilweise als Weiterentwicklung von Verfahren aus dem Vorgängerprojekt AVATecH [1, 2])
  - Benutzerfreundliche Integration dieser Verfahren in das Multimedia-Annotations- und Analyse-Tool ELAN [3]
- **Wissenschaftliche Ziele des AUVIS-Projekts**
  - Erprobung der technischen Verfahren anhand einer konkreten wissenschaftlichen Fragestellung zur sprachlichen (syntaktischen und prosodischen) und gestischen Segmentierung von (komplexen) Ereignissen
  - Untersuchung des Zusammenhangs zwischen Sprache und Gestik
  - Evaluation der Verfahren zur automatischen Analyse von Audio- und Video-Daten auf der Basis manuell annotierter Daten

Wissenschaftliche Fragestellung

- Ereignissachverhalte (z.B. das Zerbrechen einer Vase) können sprachlich auf sehr unterschiedliche Weisen kodiert werden.
  a) *Peter zerbrach die Vase.*
  b) *Peter ließ die Vase fallen und sie zerbrach.*
  c) *Die Vase entglitt Peters Händen und zerbrach.*
  d) *Peter zerbrach die Vase, indem er sie fallen ließ.*

Abb. 1: Zerbrochene Vase

Tanakawho

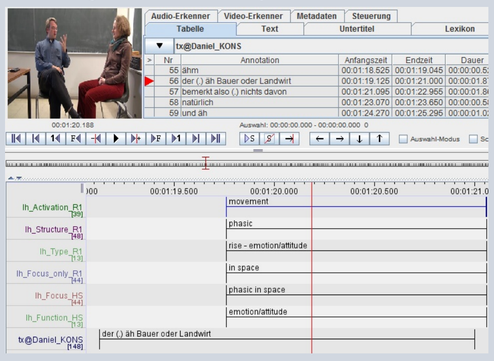
- In vielen Sprachen werden (komplexe) Ereignisse durch Konstruktionen mit mehreren Verben (sog. Verbserialisierung) ausgedrückt, z.B. in der austronesischen Sprache Wooi.

teveri	*ma*	*o:*	***mehoy***	***riapa***	*tiang*	*vaw*
3s:return	VEN	INTJ	3s:stay	3s:roast	fish	DEM

„*Er kam zurück (und) er briet den Fisch.*“ (mob1_ew 082)

- Dabei ist nicht immer klar, ob sprachlich ein (komplexes) Ereignis oder mehrere Ereignisse kodiert werden und wo die Grenzen zwischen Ereignissen / Ereignisbeschreibungen liegen.
- Manuelle Annotation von Daten aus DoBeS-Sprachdokumentationsprojekten [4, 5] und von elizitierten narrativen Texten (Pear Stories) auf syntaktischer, prosodischer und gestischer Ebene (Annotation der Gestik auf Basis des NEUROGES-Annotationsschemas [6])
- Durchführung von Interrater-Agreement-Studien zur Beurteilung der Zuverlässigkeit der Annotation
- Untersuchung möglicher Korrelationen zwischen syntaktischen, prosodischen und gestischen Einheiten in Multiverb-Konstruktionen

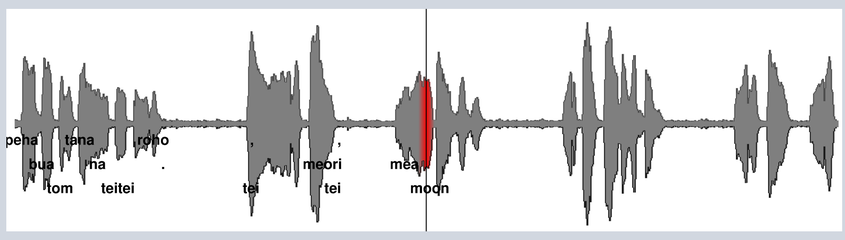
Abb. 2: Manuelle Annotation auf gestischer und sprachlicher Ebene



Automatische Audio-Analyse

- **Aufgabenstellungen**
  - Automatische Alinierung von Transkription und Audioaufnahme
  - Query-by-Example: Markieren von sprachlichen Mustern im Audio-Signal und Suche nach weiteren vergleichbaren Mustern kombiniert mit schnellen und interaktiv überwachten maschinellen Lernverfahren
    - Auf segmentaler Ebene: Suche nach Silben, Morphemen, einzelnen Wörtern oder Wortsequenzen
    - Auf prosodischer Ebene: Suche nach Intonationsmustern auf der Basis suprasegmentaler Parameter (Intensität, Längenvariation, Intonationskontur)
- **Lösungsansätze**
  - Automatische Alinierung von Transkription und Audioaufnahme mittels sprachunabhängiger Erkennung von Strukturen (z.B. Wiederholungen) im Audiosignal (vgl. Abb. 3)
  - Suche nach segmentalen und/oder prosodischen Mustern in nicht annotierten Daten („Speech recognition as a retrieval problem“) (evtl. mit vorhergehender Indizierung)
  - Automatische Segmentierung von Audiodaten in kurze phonetisch homogene Segmente mit Hilfe von Selbstähnlichkeitsmatrizen und Random Walks
  - Automatische Segmentierung in Silben

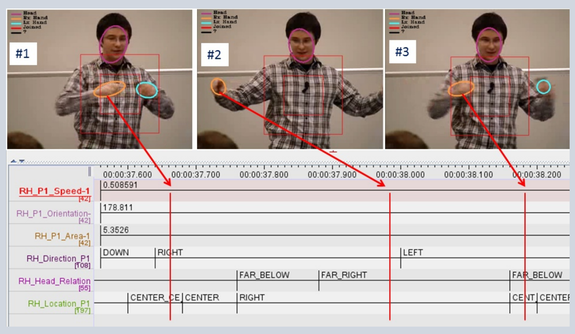
Abb. 3: Automatische Strukturerkennung in Audioaufnahmen



Automatische Video-Analyse

- **Aufgabenstellungen**
  - Automatische Erkennung von gestischen Sequenzen in Videodaten
  - Genaue Verfolgung der Position der Hände im Gestenraum („Hand-Tracking“)
  - Automatische Segmentierung und Annotation der erkannten Gesten auf der Basis des NEUROGES-Annotationsschemas [6]
  - Query-by-Example auf gestischer Ebene: Markieren einer gestischen Einheit und Suche nach weiteren vergleichbaren gestischen Mustern kombiniert mit schnellen und interaktiv überwachten maschinellen Lernverfahren
- **Lösungsansätze**
  - Automatische Verfolgung der Position der Hände auf der Basis von Hautfarbe, Bewegung sowie Hintergrundsegmentierung
  - Berücksichtigung von Kamerabewegung, bewegtem Hintergrund, Änderung der Lichtverhältnisse, Anzahl der Personen in der Szene sowie deren Bewegung
  - Automatische Annotation unterschiedlichster Eigenschaften von Händen und Gesicht, wie z.B. absolute und relative Position, Geschwindigkeit, Bewegungsbereich
  - Formulierung von Richtlinien für Videoaufnahmen für die Feldforschung

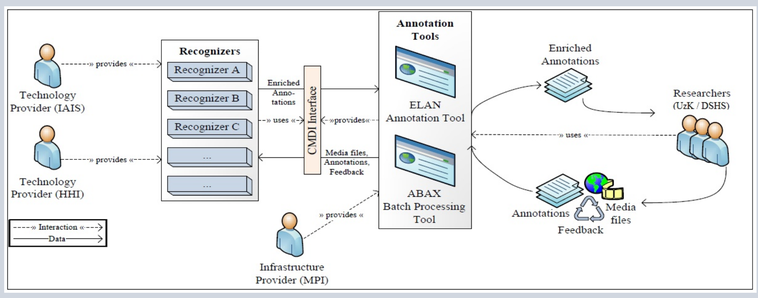
Abb. 4: Automatische Annotation von Gesten auf der Basis von Hand-Tracking-Daten



Integration der Verfahren in ELAN

- Integration der entwickelten Verfahren für die Audio- und Videoanalyse in ELAN [3]
- ELAN ist ein vielseitiges am Max-Planck-Institut für Psycholinguistik entwickeltes Multimedia-Annotations-Tool (u.a. für die Linguistik und die Gestik-Forschung) (CLARIN-D-Standard [7])
- ELAN erlaubt die Einbindung externer Anwendungen (z.B. Tools für die automatische Audio- oder Video-Analyse) und den Import der von diesen erzeugten Ergebnisse und Annotationen
  - Lokale Einbindung (Datenaustausch per XML oder CSV)
  - Einbindung als REST-Web-Service (Verwendung von CLAM als Wrapper)
  - Automatische Erzeugung einer passenden GUI auf der Basis von CMDI-Metadaten [8]
  - Einfache Feedback- und Korrekturmechanismen für den Benutzer

Abb. 5: ELAN als Arbeitsumgebung für automatisierte Annotation



Referenzen

[1] Fraunhofer - Max Planck Forschungsprojekt AVATecH: http://www.mpi.nl/avatech
[2] Schreer, Oliver und Schneider, Daniel (2012). Supporting linguistic research using generic automatic audio/video analysis. In: *Potentials of Language Documentation: Methods, Analyses, and Utilization*, hg. v. Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts und Paul Trilsbeek, 39–46.
[3] ELAN: http://tla.mpi.nl/tools/tla-tools/elan/
[4] DoBeS: http://www.mpi.nl/DOBES/
[5] The Language Archive: http://tla.mpi.nl/
[6] Lausberg, Hedda und Sloetjes, Han (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41(3), 841-849.
[7] CLARIN-D: http://de.clarin.eu/de/sprachressourcen/werkzeuge-und-dienste.html
[8] CMDI (Component MetaData Infrastructure): http://www.clarin.eu/cmdi

Kontakt / Projektpartner

BMBF-gefördertes Verbundforschungsprojekt in den eHumanities.
Förderzeitraum: 07/2012 - 06/2015

Projekt-Webseite: http://tla.mpi.nl/projects_info/auvis/

Universität zu Köln
Philosophische Fakultät
Institut für Linguistik
Lehrstuhl für Allgemeine Sprachwissenschaft
Prof. Dr. Nikolaus Himmelmann

Deutsche Sporthochschule Köln
Abteilung für Neurologie, Psychosomatik und Psychiatrie
Institut für Bewegungstherapie und bewegungsorientierte Prävention und Rehabilitation
Prof. Dr. Hedda Lausberg

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin
Speech Technology Group
Dr. Rolf Bardeli

Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut, Berlin
Image Processing Department
PD Dr. Oliver Schreer

Max-Planck-Institut für Psycholinguistik, Nijmegen
The Language Archive
PD Dr. Sebastian Drude

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung