Das Korpus der kroatischen Sprache "Hrvatska jezična mrežna riznica"

Dunja Brozović Rončević & Damir Ćavar

riznica@ihjj.hr

Institut za hrvatski jezik i jezikoslovlje

19. Juni 2006

© 2006 Institut za hrvatski jezik i jezikoslovlje

Agenda

- **Das Korpus der kroatischen Sprache**
  - Hintergrund
  - Ziele
  - Konzeption und Realisierung
  - Technische Details
- **Forschungsthemen**

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektstruktur

- **Institut za hrvatski jezik i jezikoslovlje** (*Institut für kroatische Sprache und Sprachwissenschaft*)
  - Dr. Dunja Brozović Rončević
  - Dr. Małgorzata E. Ćavar
  - Dr. Damir Ćavar
  - Ivo-Pavao Jazbec
  - Dr. Željko Jozić
  - Mag. Tomislav Stojanov
  - und viele weitere Mitarbeiter des Instituts

© 2006 Institut za hrvatski jezik i jezikoslovlje

Hintergrund

- Beantragt beim Ministerium für Wissenschaft, Bildung und Sport der Republik Kroatien Ende 2004
- Genehmigt im Mai 2006, als Forschungsprojekt am Institut für kroatische Sprache und Sprachwissenschaft
- Projektleitung: Dr. Dunja Brozović Rončević
- Leitung der Sektion theoretische und Computerlinguistik: Dr. Damir Ćavar

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektziele

- **Schaffung eines Internet-Archivs für Kroatisch**
  - Korpora in verschiedenen Varianten
  - Wörterbücher, und insbesondere Fachwörterbücher
  - Online-Berater und Informationsseiten für Standard-Kroatisch
  - Digitale Landkarten mit Texten und Aufnahmen kroatischer Mundarten

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektziele

- **Schaffung eines Internet-Archivs für Kroatisch**
  - Digitale Ressourcen für Sprachattechnologie, e. g. statistische Modelle (N-Gram Modelle), Wortlisten, Aufnahmen, Grammatiken
  - Grundlegende Sprachtechnologien, i. e. Tagger oder Part-of-Speech Annotationsprogramme, Lemmatisierer, Parser

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektziele

- **Zielgruppe**
  - Sprach- und Literaturwissenschaftler, z. B. im Bereich der Kroatistik und verwandter komparativer Forschung
  - Ausschuss für die Normierung der kroatischen Standardsprache des Ministeriums für Wissenschaft, Bildung und Sport der Republik Kroatien
  - Weitere nationale und europäische Institutionen

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektziele

- **Zielgruppe**
  - Die Öffentlichkeit, Bildung und Medien
  - Wirtschaft, und Forschung und Entwicklung im Bereich Sprachtechnologie

© 2006 Institut za hrvatski jezik i jezikoslovlje

Ausgangspunkt

- Zentrale Verknüpfung existierender verwandter Projekte und Aktivitäten, u. a. auch zur Maximierung der Synergie und Minimierung von Redundanzen
- Forschung am Kroatischen diachron und synchron
  - nicht allein am Standardkroatischen, sondern an einer eigenen und gemeinsamen kroatischen kulturellen Gesamtheit orientiert

© 2006 Institut za hrvatski jezik i jezikoslovlje

Projektprinzipien

- Technologien und Verfahren, wie auch Resultate:
  - Offene Standards
  - Frei verfügbare Ressourcen

© 2006 Institut za hrvatski jezik i jezikoslovlje

Korpus der kroatischen Sprache

- **Kein Einzelkorpus sondern eine Gruppe von Korpora**
  - Korpus des Altkroatischen (bis zur Mitte des 15. Jh.)
  - Korpus des Mittelkroatischen (bis zum Beginn der Standardisierung, d. h. Mitte des 18. Jh.)
  - Korpus der kajkavischen Literatursprache
  - Standardkroatischer Korpus (vom Beginn der Standardisierung bis heute)

© 2006 Institut za hrvatski jezik i jezikoslovlje

Korpus der kroatischen Sprache

- **Momentane Arbeitsphase**
  - Altkroatisch: Ablichtung
  - Mittelkroatisch: Ablichtung
  - Kajkavische Literatursprache: Ablichtung und Digitalisierung
  - Mundarten: Digitalisierung und Transkription von Audiomaterial
  - Standardkroatisch: Teile online verfügbar
    - Belletristik
    - Fach- und Sachbücher
    - Zeitungstexte, usw.

© 2006 Institut za hrvatski jezik i jezikoslovlje

Korpus der kroatischen Sprache

- **Standardkroatisch: Auswahlprinzip**
  - Alle Texte werden in ihrer ursprünglichen Form ausgewählt, d. h. in der Form wie sie vom Autor zur Publikation freigegeben wurden.
  - Diese Ausgaben werden durch Ausgaben aus Zeiten der jugoslawischen Diktatur komplementiert, um eine komparative Analyse der Variation zu ermöglichen.

© 2006 Institut za hrvatski jezik i jezikoslovlje

Korpus der kroatischen Sprache

- Textquellen
  - Ablichtungen und Digitalisierungen von Erstausgaben
  - Digitale Ausgaben der Verlage
    - Školska knjiga
    - Stoljeća hrvatske književnosti, Matica hrvatska
    - und weitere Verträge in Vorbereitung

© 2006 Institut za hrvatski jezik i jezikoslovlje

Korpus der kroatischen Sprache

- Online-Zugang zum Korpus: riznica.ihjj.hr
  - Volltext-Suche
  - Statistische Analysen und Daten (z. B. Wortfrequenzen, Kollokationen, Schlagwörter im Kontext)
- In Vorbereitung:
  - Zugang zu den abgelichteten Originalen
  - Suche über linguistische Funktionen und Strukturen

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- Ablichtung
- Digitalisierung (Konvertierung in ein Unicode-Textformat, Audio-Digitalisierung)
- Annotation
- Qualitätsprüfung
- Indexierung

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- **Ablichtung**
  - Scanning und OCR
  - Erzeugung von Bildformaten (e. g. TIFF)
  - Erzeugung von PDF Versionen
  - Erzeugung von Meta-Dateien
    - Titel, Autor, Ausgabedetails, Arbeitsschritte und Verantwortliche

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- **Digitalisierung von Texten**
  - Unicode Kodierung, für unsere Zwecke UTF-8
  - Erzeugung von:
    * PDF- und RTF-Dateien mit Formatinformationen
    * Rohtext-Dateien ohne Formate
    * Vorverarbeitung von Formatinformationen
- Erzeugung roher XML-Dateien

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- **Annotation:** Standard festgelegt durch die *Text Encoding Initiative* (TEI) URL: www.tei-c.org
- eXtensible Markup Language (XML) Format
- Standardisierte Meta-Informationstypen und Strukturelemente
- Offen für spezifische Anpassungen und Erweiterungen

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- XML und TEI XML
  - Offener und nicht proprietärer Standard
  - Frei verfügbare Programme für die Verarbeitung, Konvertierung, Analyse
  - Einfache Übertragung auf professionelle Datenbanken
- Vorteile durch professionelle Vorarbeiten von TEI Experten

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- **Stufenweise Annotation:**
  - Meta-Information
  - Formatinformation, e. g. Kursiv, Fett
  - Textstruktur
    - e. g. Buchteile, Kapitel, Artikel
    - Paragraphen, Sätze, Wörter
  - Linguistische Annotation

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- **Linguistische Annotation in TEI XML vorgesehen**
  - Lemma
  - Part-of-Speech, d. h. morpho-syntaktische Annotation
- aber offen für Erweiterungen und Spezifikationen, z. B.
  - Semantische Merkmale
  - Spezifische Auswahl der *Part-of-Speech Tags*

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- Vorteile der partiellen Annotationsmöglichkeit
  - Schrittweiser Ausbau des Korpus in die Tiefe, bez. inhärenter textueller Merkmale oder linguistischer Ebenen
  - Aufteilung der Aufgaben in Unteraufgaben und Zuweisung von Experten
  - Spezifische Qualitätskontrolle nach jedem Arbeitsschritt

© 2006 Institut za hrvatski jezik i jezikoslovlje

Arbeitsschritte

- Vorteile der partiellen Annotation
  - Kombinierbar mit Versionierungssystemen und Parallelverarbeitung
  - Unmittelbare Integration der Zwischenergebnisse in eine Arbeitsumgebung

© 2006 Institut za hrvatski jezik i jezikoslovlje

Schnittstelle

- Philologic (University of Chicago)
  - URL: philologic.uchicago.edu
  - Schlagwortsuche und Kontext
  - Einfache Statistik
  - Kollokationsanalyse
  - Einschränkung auf Meta-Informationen wie z. B. Zeitachse, Autor, Ort

© 2006 Institut za hrvatski jezik i jezikoslovlje

Schnittstelle

- Philologic Vorteile
  - Offene und freie Programm-Quellen
  - Zugang zum Korpus über das Internet mit beliebigen Web-Browsern
  - Offline Verarbeitung des Korpus und Geschwindigkeit bei der Suche
  - Volle TEI XML Unterstützung und Unicode Zeichen-Kodierung

© 2006 Institut za hrvatski jezik i jezikoslovlje

Schnittstelle

- Philologic Nachteile
  - Server-basiertes System, das hochwertige Technik und Expertenwissen für die Installation, Anpassung und Erweiterung verlangt
  - Offline Verarbeitung des Korpus und Geschwindigkeit bei der Indexierung (Prozessor-lastig, zeit- und speicheraufwendig)
  - Eingeschränkte Analyse und Arbeitswerkzeuge

© 2006 Institut za hrvatski jezik i jezikoslovlje

Schnittstelle

- Beispiel online
- Unsere Vorgehensweise
  - Erweiterung durch eigene Softwareentwicklung: z. B. Funktionalität, selektive Beschränkungen, Optimierung des Arbeitsprozesses

© 2006 Institut za hrvatski jezik i jezikoslovlje

Momentaner Stand

- Annotationstiefe:
  - Textstruktur bis auf Paragraph-Ebene
- Tiefere Annotation semiautomatisch:
  - Linguistische Analysekomponenten in Entwicklung: Satzerkennung, morph-syntaktische Annotation, syntaktischer Parser mit semantischer Struktur, $Named\ Entity$ Erkenner

© 2006 Institut za hrvatski jezik i jezikoslovlje

Momentaner Stand

- Quantitativer Stand
  - über 71 Mil. Wörter ($Tokens$)
  - über 650.000 unterschiedliche Wörter ($Types$)

© 2006 Institut za hrvatski jezik i jezikoslovlje

Weitere Planung

- Analyse zeitlicher Entwicklung quantitativer lexikalischer Eigenschaften (Uni- und N-gram Modelle über die Zeitachse)
- Kontextuelle Eigenschaften lexikalischer Modelle anhand von Vektor-Raum Modellen (lexikologische Analyse)
- Generierung semantischer Netze (über linguistische Annotationen)

© 2006 Institut za hrvatski jezik i jezikoslovlje

Weitere Planung

- Distanz- und Korrelationsanalysen anhand tiefer linguistischer Annotation
- Bereitstellung grundlegender Sprachtechnologien (Lemmatisierer, $Tagger$, Parser)

© 2006 Institut za hrvatski jezik i jezikoslovlje

Weitere Planung

- Daten für Sprachtechnologien
  - Trainingskorpora
  - Lexika,
  - N-Gram Modelle auf der phonologischen, orthographischen, lexikalischen, syntaktischen Ebene

© 2006 Institut za hrvatski jezik i jezikoslovlje