

Soziolinguistik trifft Korpuslinguistik

Christoph Draxler
Bayerisches Archiv für Sprachsignale BAS
draxler@phonetik.uni-muenchen.de

Danke an das CLARIN- und Text+-Team in München

<https://clarin.phonetik.uni-muenchen.de/BASRepository/>

Soziolinguistik trifft Korpuslinguistik



Herausgegeben von
Björn Hansen
Anna Zielińska

Open Access Publikation
<https://doi.org/10.33675/2019-82538591>

LangGener-Projekt

Zwei Forschungsgruppen, drei Teams, ein Ziel:

Deutsch-polnischer Bilingualismus in Deutschland und Polen

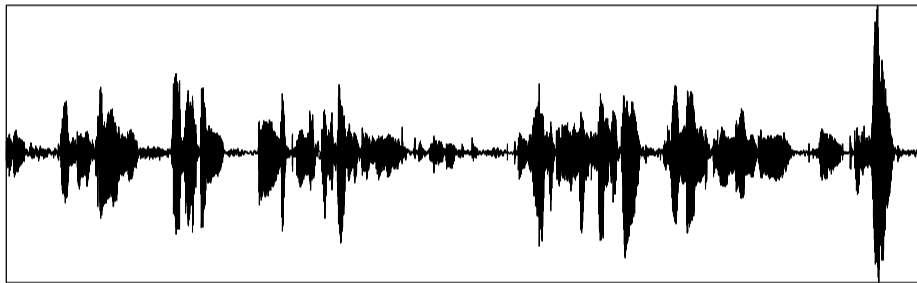
- ▶ Institut für Slavistik, Universität Regensburg, Björn Hansen
- ▶ Institut für Slavistik der Polnischen Akademie der Wissenschaften, Anna Zielińska
- ▶ Center for Czech Studies, Marek Nekula (tschechisches Subkorpus)

BEETHOVEN Förderinitiative für integrierte akademische deutsch-polnische Forschungsprojekte

Sprachaufnahmen im Feld

- ▶ Arbeitsplan: je 30 Personen in Deutschland und Polen aufnehmen
- ▶ Sprachkontaktphänomene sowie Dialektmerkmale
- ▶ tatsächlich 124 Personen aufgenommen
 - ▶ 47 'Generation Deutschland'
 - ▶ 77 'Generation Polen'
- ▶ im Korpus: 58 bilinguale Personen mit insgesamt 78 Stunden Gesprächsaufnahmen

Beispielaufnahme (aus dem tschechischen Subkorpus)



war für mich alleine oder im im vikýř das weiß ich nicht wie man vikýř sagt sehen Sie
das ist so ein Fenster wo man wo man die ne das Heu

Interviews

- ▶ Rekrutierung durch Einladungsschreiben an Institutionen, Zeitungen und persönliche Kontakte
- ▶ Interviews meist zuhause
- ▶ Einwilligungserklärung für die Aufnahme sowie – pseudonymisiert – die wissenschaftliche Analyse und Nachnutzung
- ▶ teilnehmende Beobachtung, sichtbare Aufnahmetechnik
- ▶ Ansprache auf deutsch und polnisch
- ▶ Dankesbrief nach jedem Interview

Transkription

Transkriptionskonventionen angepasst an die Forschungsziele

- ▶ Dokumentation der Dialekte
- ▶ Typen morphosyntaktischer Replikation
- ▶ Analyse der Sprachbiographien
- ▶ Beziehung zwischen Sprachbiographie und Replikation
- ▶ Korrelationen der Typen von Replikationen

Transkription

Transkriptionskonventionen angepasst an die Forschungsziele

- ▶ Dokumentation der Dialekte
- ▶ Typen morphosyntaktischer Replikation
- ▶ Analyse der Sprachbiographien
- ▶ Beziehung zwischen Sprachbiographie und Replikation
- ▶ Korrelationen der Typen von Replikationen

Die standardnahe orthographische Transkription (deutsch, polnisch und tschechisch) erleichtert das Auffinden von Konstruktionen und 'schließt somit etwa mehrere phonetische Varianten einer Wortform ein'

Transkription

Transkriptionskonventionen angepasst an die Forschungsziele

- ▶ Dokumentation der Dialekte
- ▶ Typen morphosyntaktischer Replikation
- ▶ Analyse der Sprachbiographien
- ▶ Beziehung zwischen Sprachbiographie und Replikation
- ▶ Korrelationen der Typen von Replikationen

Die standardnahe orthographische Transkription (deutsch, polnisch und tschechisch) erleichtert das Auffinden von Konstruktionen und 'schließt somit etwa mehrere phonetische Varianten einer Wortform ein'

Segmente umfassen grundsätzlich ein konjugiertes Verb, einschließlich Interjektione und Häsitationen.

Transkription

Transkriptionskonventionen angepasst an die Forschungsziele

- ▶ Dokumentation der Dialekte
- ▶ Typen morphosyntaktischer Replikation
- ▶ Analyse der Sprachbiographien
- ▶ Beziehung zwischen Sprachbiographie und Replikation
- ▶ Korrelationen der Typen von Replikationen

Die standardnahe orthographische Transkription (deutsch, polnisch und tschechisch) erleichtert das Auffinden von Konstruktionen und 'schließt somit etwa mehrere phonetische Varianten einer Wortform ein'

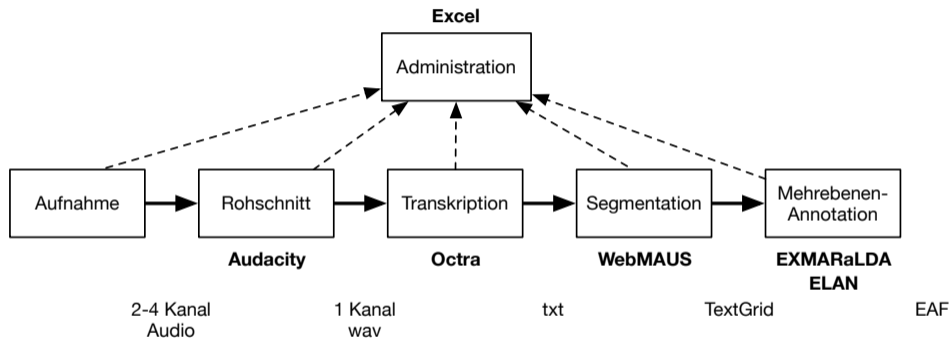
Segmente umfassen grundsätzlich ein konjugiertes Verb, einschließlich Interjektione und Häsitationen.

Marker und Symbole wie ", '#', unverständliche Passagen in . . . , keine Satzzeichen außer '?'

Vom Gespräch zu annotierten Daten

- ▶ Aufnahme mit Headset oder Krawattenmikrofon – gute Signalqualität erleichtert die Transkription!
- ▶ manuelle Bestimmung von relevanten Interview-Abschnitten
- ▶ systematische Benennung aller Dateien, z. B. KG_MUN_MI_DE_0000
- ▶ manuelles Schneiden in Audacity, erfassen aller Dateien in einer Masterliste in Excel
- ▶ manuelle Segmentation und Transkription sowie automatisches Schneiden in Ocrata

Verarbeitungs-Workflow



Der Workflow basiert auf Tools und Webdiensten aus Text+

Metadaten

Von Beginn an systematische Erhebung und Erfassung der Metadaten in Excel-Tabellen

speakers personenbezogene Angaben

interviewers Angaben zu Interviewern

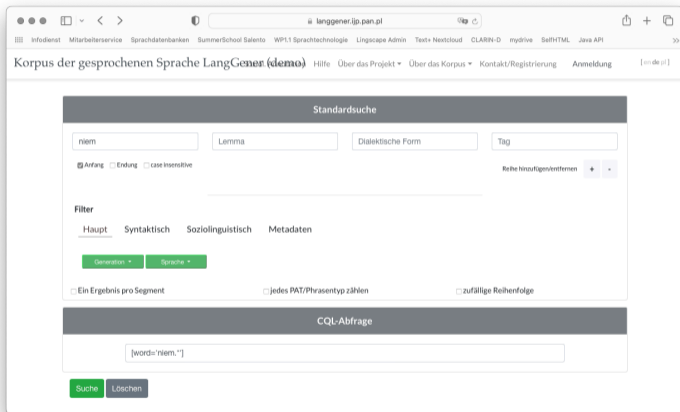
interviews Aufnahmesituation, Sprache, Speicherort, usw.

files alle Dateien eines Interviews

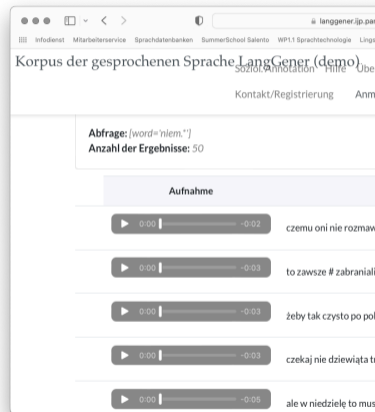
Die Endkontrolle aller Dateien im Korpus erfolgte über eine interaktive web-basierte Liste gemeinsam durch beide Forschungsgruppen

Verfügbarkeit und Nachnutzung

Das LangGener-Korpus ist über die Polnische Akademie der Wissenschaften nutzbar:



The screenshot shows the 'Standardsuche' (Standard Search) section of the LangGener website. It features several input fields for search criteria: 'niem' in the main search box, and 'Lemma', 'Dialektische Form', and 'Tag' in smaller boxes. Below these are checkboxes for 'Anfang', 'Endung', and 'case insensitive', along with a 'Reihe hinzufügen/entfernen' button. A 'Filter' section includes tabs for 'Haupt', 'Syntaktisch', 'Soziolinguistisch', and 'Metadaten'. There are also dropdown menus for 'Generation' and 'Sprache'. At the bottom of this section are checkboxes for 'Ein Ergebnis pro Segment', 'Jedes PAT/Phrasentyp zählen', and 'zufällige Reihenfolge'. Below the search section is a 'CQL-Abfrage' (CQL Query) section with a text input containing '[word="niem.*"]' and 'Suche' and 'Löschen' buttons.



The screenshot shows the search results page for the query '[word="niem.*"]'. It displays 'Anzahl der Ergebnisse: 50'. Below this is a section titled 'Aufnahme' (Recording) which lists several audio clips. Each clip has a play button, a progress bar, and a duration. The clips are: 'czemu oni nie rozmaw...', 'to zawsze # zabraniali...', 'żeby tak czysto po pol...', 'czekaj nie dziewiąta t...', and 'ale w niedzielę to mus...'. The website header includes navigation links like 'Hilfe', 'Über das Projekt', 'Über das Korpus', 'Kontakt/Registrierung', and 'Anmeldung'.

<https://langgener.ijp.pan.pl>

Archivierung

- ▶ Das LangGener Korpus wird nun in das Repository des BAS importiert – die Arbeitsgruppe ist im Aufbau
- ▶ Der Webservice COALA erstellt aus den Metadaten-Tabellen CMDI-konforme Metadaten für das Repository

Web-Dienste und Werkzeuge für die Sprachverarbeitung

Einfacher Zugang zu aktueller Sprachtechnologie

- ▶ für Forscher/innen, Entwickler/innen und Studierende
- ▶ mit dem Fokus auf gesprochene Sprache
- ▶ in vielen Sprachen
- ▶ und unterschiedlichen Forschungs- und Anwendungsgebieten

Die Nutzung der Dienste ist für akademische Nutzer/innen gratis.

Browser tabs: BAS | web service interface | 1: X

Address bar: clarin.phonetik.uni-muenchen.de/BASWebServices/interface

Header: **BAS Web Services** Version 3.1 · History of changes

Navigation: Home | General Help + FAQs | Publications | Contact/About

Show service sidebar >

Welcome! The BAS Web Services are a rich set of tools for speech sciences and technology. Starting with MAUS – automatic segmentation and labelling of speech – many tools were developed in the context of [CLARIN-D](#).

Developing these tools is scientific work. Please cite the tools in your publications – every tool comes with a reference, and there is a [publications page](#).

For further information, there is a [General Help](#) page, and for every service there is a specific help section. All tools can be used via this graphical frontend, or via a programmatic REST API.

By using the services, you accept the [Conditions of Use](#) of these services.

ASR
Automatic transcription of speech signal using several third party ASR services (experimental).
This service requires authentication and is only free for academic users.

Anonymizer
This services reads a signal file (sound, video) + BAS Partitur Format annotation + a list of terms to be anonymized in both inputs, masks all occurrences in the signal and in the annotation, and returns the two anonymized files in a ZIP archive.

AudioEnhance
This services reads a media file and performs several signal processing operations mostly

Formant Analysis

CLARIN-D INSTITUTE OF PHONETICS AND SPEECH PROCESSING

IPS

Berlinian Archive for Speech Signals