Entwicklung einer vergleichbaren Multi-Ebenen-Annotation von gesprochenen Sprachkontaktdaten

Margaret Blevins
Department of Germanic Studies
mblevins@utexas.edu

FRAGESTELLUNG

MOTIVATION

Obwohl sich heutzutage viele Forschungsprojekte mit der Dokumentation deutschsprachiger Sprachinseln bzw. Sprachkontaktvariatäten beschäftigen, fällt auf, dass die entstandenen Korpora insgesamt hinsichtlich der Erhebungs-, Transkriptions- und Annotationsmethoden schwer miteinander zu vergleichen sind.

Um eine systematisch vergleichende Sprachkontakt- bzw. Sprachinselforschung zu ermöglichen, müssen passende Daten für vergleichende Fragestellungen bereitgestellt werden, z.B. durch systematische Transkription und Annotation (vgl. Boas 2016: 38-40).

ZIELSETZUNG

Es ist Ziel meines Dissertationsprojektes, die Vergleichbarkeit von Sprachinseldaten zu verbessern. Dies soll durch die Entwicklung eines einheitlichen Multi-Ebenen Annotationssystem für gesprochene Sprachkontaktdaten realisiert werden.

HERAUSFORDERUNGEN

Vergleichende Sprachkontaktforschung ist aus zwei Gründen besonders komplex: 1) Variation innerhalb einzelner Korpora und 2) Variation zwischen verschiedenen Korpora.

Möchte man mit mehreren Korpora gleichzeitig arbeiten, ist es besonders schwer, vergleichbare Informationen zu finden bzw. Schlussfolgerungen zu ziehen.

I. Variation innerhalb eines Korpus

Gesprochene Sprache → **Aufnahme** → **Transkription**

Variationsquellen

Gesprochene Sprache	Aufnahme	Transkription
• Inter- bzw. Intra-Sprecher	• Rauschen	• Inter-Transkribant
• Dialekte	• Mehrere Sprecher	• Intra-Transkribant
• Sprachkontaktphänomene (z.B. Codemixing)	• Lautstärke	• Tippfehler
• Disfluenzen (z.B. Füllwörter, Wortwiederholungen)	• Aufnahmegerät	• Missverständnisse

II. Variation innerhalb mehrerer Korpora

Variationsquellen innerhalb mehrerer Korpora:

- unterschiedliche **Kontaktsprachen** (z.B. Elsässisch-Englisch vs. Schwäbisch-Russisch)
- unterschiedliche **Fragestellungen** bzw. Forschungsinteressen, Fachkenntnisse, Unterstützung (technisch, finanziell, Hilfskräfte)
- unterschiedliche **Erhebungsmethoden** (z.B. freie Gespräche, Übersetzungen, Sprachsituationen, usw.)
- verschiedene **Annotationsebenen** & **Dateiformate**

Was annotiert wird	Sprachkontaktkorpora: Unserdeutsch	RuDiDat	TGDP	UZSK	DNam
Transkription (phonetisch)	IPA	IPA			
Transkription (orthographisch)	cGAT			GAT2 (modifizierte minimal Transkription)	cGAT (minimal Trankription)
Normalisierung					Duden & FOLK
Lemma					(TreeTagger)
Wortart	STTS 2.0 (leicht modifiziert)	STTS (reduziert & modifiziert)			STTS 2.0 (+Ergänzungen)
Sprachkontakt Phänomene		+ 1:1 ins Deutsch übersetzt	+ 1:1 ins Englische übersetzt	+ ins Deutsch übersetzt	
Übersetzung					

Tab. 1. Übersicht von Variation in unterschiedlichen Sprachkontaktkorpora

= Annotationsebenen, die nach systematischen, **NICHT- projektbezogenen** Richtlinien erstellt wurden

= Annotationsebenen, die nach **projektbezogenen** Richtlinien erstellt wurden

RuDiDat = Russlanddeutsche Dialektdatebank
TGDP = Texas German Dialect Project
UZSK = Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus
DNam = Deutsch in Namibia

METHODE

Auf der Basis der Daten des Texas-German-Dialect-Projects (TGDP, Boas et al. 2010) wird ein einheitliches Annotationssystem entwickelt, das auf verschiedene Sprachinsel- und Sprachkontaktkorpora allgemein anwendbar ist.

DATEN

Laut Boas (2009:34) ist **Texasdeutsch** „a set of varieties of German spoken in Texas which have descended from the dialects of German brought to Texas in the 19th century.“

Das Konzept basiert auf der Auswertung eines kleinen, repräsentativen Korpuses bestehend aus TGDP **Übersetzungsdaten.** Sprecher des Texasdeutschen übersetzen eine Liste Englischer Sätze und Wörter ins Texasdeutsche (vgl. Gilbert 1972). Die Übersetzungen von **30 randomisiert ausgewählten SprecherInnen** werden anschließend auf verschiedenen Ebenen annotiert.

TOOLS

Die Daten wurden mit dem EXMARaLDA Partitur-Editor annotiert (Schmidt 2016).

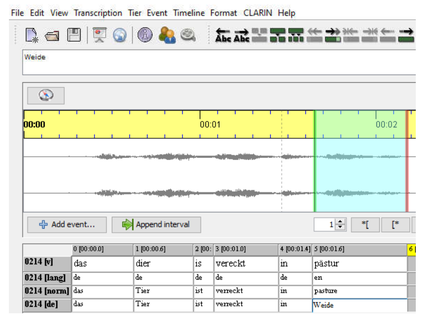

Abb. 1. Screenshot des EXMARaLDA Partitur-Editors

ANFORDERUNGEN & VORAUSSETZUNGEN FÜR DIE ANNOTATION

- klare, verständliche Richtlinien, die für alle zugänglich sind → transparent & reproduzierbar
- relativ flexibel & gleichzeitig konsistent
- kompatibel mit mehreren Forschungsparadigmen

VORLÄUFIGES SYSTEM

ZIELE

Hauptziel: Vergleichbarkeit der Daten. Zu diesem Zweck gilt es, eine Annotationsebene zu etablieren, auf der Variation systematisch reduziert ist (z.B. Schreibvarianten eines „Wortes“). Es ist *nicht* das Ziel, die Texasdeutschen Texte zu ‚korrigieren‘ oder komplett in den Standard zu ‚übersetzen‘.

ANNOTATIONEN: MULTI-EBENEN-STRUKTUR

Sprach-Tagging (lang)

- basiert auf Phonologie & Morphologie
- 6 Hauptkategorien (Tags)
  - de (Deutsch)
  - en (Englisch)
  - lang3 (andere Sprache)
  - mix (z.B. *gejumpt*)
  - ambig
  - unknown

Normalisierung (norm)

- Lemma & Wortart der originalen Token erhalten
- die Normalisierung variiert je nach Sprach-Tag
- **nicht** verändert werden: Genus, Kasus, Tempus, Aspekt, Modus, Semantische Wortwahl, Wortstellung
- keine Tokens hinzufügen oder löschen

Minimale Übersetzung (de)

- alle Tokens, die nicht mit „de“ getaggt sind, werde ins Deutsche übersetzt
- **nicht** verändert werden: Genus, Kasus, Tempus, Aspekt, Modus, Semantische Wortwahl, Wortstellung
- keine Tokens hinzufügen oder löschen

VORTEILE DES SYSTEMS

Sprach-Tagging (lang)

- Sprachkontaktphänomene sind auffindbar, ohne sie von vornherein zu kategorisieren
- Sprachzuordnung ist für BenutzerInnen nachvollziehbar
- Tags können hierarchisch erweitert werden (z.B. de-b für Bairisch)

Normalisierung (norm)

- Flexibilität ermöglicht mehrere Sprachen bzw. Dialekte zu annotieren
- zu komplexe ‘Korrekturen’/ Interpretationen werden vermieden (→ einheitlicher / konsistenter)

Minimale Übersetzung (de)

- alles in einer Sprache (→ einfacher zu vergleichen)

Bitte beachten: Das oben dargestellte System ist nicht für *alle* Fragestellungen perfekt geeignet. Viele andere Ebenen können bzw. sollen hinzugefügt werden, z.B. phonetische Transkriptionen oder Zielhypothesen.

QUELLEN

Boas, Hans C. 2009. *The life and death of Texas German*. Durham: Duke University Press.

Boas, Hans C., Marc Pierce, Karen Roesch, Guido Halder, & Hunter Weilbacher. 2010. The Texas German Dialect Archive: A multimedia resource for research, teaching, and outreach. *Journal of Germanic Linguistics* 22 (3), 277-296.

Götze, Angelika, Siegwalt Lindenfelser, Salome Lipfert, Katharina Neumeier, Werner König, & Péter Maitz. 2017. Documenting Unserdeutsch (Rabaul Creole German): A workshop report. *Journal of the Linguistic Society of Papua New Guinea*. 65–90.

Maitz, Péter, Werner König, Craig Alan Volker. 2016. Unserdeutsch (Rabaul Creole German): Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea. *Zeitschrift für germanistische Linguistik* 44 (1), 93-96.

Frick, Elena. 2019. Russlanddeutsche Dialekte. http://prowiki.ids-mannheim.de/bin/view/Russlanddeutsch/WebHome

Földes, Csaba. 2016. Ungarndeutsches Zweisprachigkeits- und Sprachkontaktkorpus: Konzept, Design und Inhalte. *Zeitschrift für interkulturelle Germanistik* 7 (1), 167-181.

Schmidt, Thomas. 2016. EXMARaLDA Partitur-Editor. Manual. <https://www.exmaralda.org/pdf/Partitur-Editor_Manual.pdf>

Zimmer, Christian, Heike Wiese, Horst J. Simon, Marrianne Zappen-Thomson, Yannic Bracke, Britta Stuhl, Thomas Schmidt (eingereicht) Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations– und Soziolinguistik.