Die orthographische Normalisierung deutschsprachiger Kontaktvarietät-Daten

Margaret Blevins
Department of Germanic Studies
mblevins@utexas.edu

FRAGESTELLUNG

Dieses Dissertationsprojekt befasst sich mit der orthographischen Normalisierung deutschsprachiger Kontaktvarietäten. Ich schlage ein Multiebenen-System zur orthographischen Normalisierung deutschsprachiger Kontaktvarietäten vor (Normalisierung + Sprach-Tagging). Meine Studie konzentriert sich vorerst auf die Daten des *Texas-German-Dialect-Projects* (Boas et al. 2010).

HERAUSFORDERUNGEN

Wie bei vielen Sprachkontaktvarietäten gibt es auch beim Texas-Deutsch zahlreiche Unterschiede im Vergleich zum (geschriebenen) Standarddeutsch. Wenn man solche Daten literarisch transkribiert (vgl. dazu Selting et al. 2009:360), fallen noch mehr Unterschiede auf. Daraus ergeben sich zahlreiche Schwierigkeiten bei der Arbeit mit den Daten: Beispielsweise erschwert die uneinheitliche Schreibung das Auffinden eines bestimmten Lexems und auch die Weiterverarbeitung durch automatische Tagger funktioniert nur eingeschränkt.

Das vorgeschlagene System setzt auf zwei Ebenen an, und zwar Anreicherung mit Metadaten (lang) und Normalisierung auf der phonologischen Ebene (norm).

= in **lang** getagt = in **norm** normalisiert

Variation innerhalb Texasdeutsch Transkriptionen

Sprachkontaktphänomene

Fremdsprachliches Lexem
- *die Kuh ist über die **fence** ge**jump**t*
  'die Kuh ist über den Zaun gesprungen'
- *Sie haben gefragt ob sie **– how do you say 'spies'** – ob sie **spies** waren*
  'Sie haben gefragt ob sie—wie sagt man 'spies' - ob sie Spione waren'

Bedeutungswandel
- *Hochschule* 'Highschool' (Englisch: high school)
- *Luftschiff* 'Flugzeug'
- *marode* & *ausgespielt* 'müde, erschöpft'

Inter-Token-Mixing
(Morpheme aus verschiedenen Sprachen)
- *fencen* 'Zäune' (fence$_{eng}$+n$_{deu}$)
- *spindlig* 'spillerig' (spindl$_{eng}$+ig$_{deu}$)
- *fortgemoved* 'fortgezogen' (fort$_{deu}$+ge$_{deu}$+move$_{eng}$+d$_{eng}$)

Lehnübersetzung (Phrasen)
- *ich kann nicht **für sicher** sagen*
  'ich kann nicht **mit Sicherheit** sagen'
- *Ich konnte nicht mit farming **ein Leben machen***
  'ich konnte nicht von der Landwirtschaft leben'

Inter– bzw. Intra-Individuelle Variation

Phonologie
- *tier* [**t**i:e̯] vs. *dier* [**d**i:e̯] 'Tier'
- *hühner* ['h**y**:ne] vs. *hiehne*['h**i**:ne] 'Hühner'

Lexikon
- *Magenschmerzen* vs. *Leibweh* vs. *Panzweh*

Morphosyntax
- Kasus: *mit **sie*** vs. *mit **ihr***
- Genus: ***das** Baum* vs. ***der** Baum*
- Progressiver Aspekt: *er **läuft*** vs. *er ist **am Laufen***
- Konjunktiv II: *ich **täte** sagen* vs. *ich **würde** sagen*
- Possessiv: *meine Frau ihre Mutter* vs. *die Mutter meiner Frau*
- Wortstellung: *ich bin gestern **gegangen** mit Mama* vs. *ich bin gestern mit Mama **gegangen***

Abbrüche
- Wort-intern
  *ja das war wo die **Ur-** auf den Platz wo der Urgroßvater [...] in **achtzehnsieben-** achtzehnsechsundsiebzig gegrind hat*
- Satz-intern
  *wie haben paar Mexikaner gehabt **because mexikaner kinder** ne Masse von die deutschen Farmers daoben haben immer ne kleines Haus gehat und da war immer ne Familie von Mexikaner drin*

Inter– bzw. Intra-Transkribant

Verwendung von Sonderzeichen
- *gro**ß*** vs. *gro**ss***
- ***ü**ber* vs. ***ue**ber*

Lage und Einbeziehung von Wortgrenzen
- *in English **nennses** 'links'* vs. *in English **nen se s** 'links'*
  'Im Englischen nennt man sie 'links' (=Verbindungen))'

Unterschiedliche Identifizierung/Interpretation
z.B. *for* vs. *vor* [fo:e̯]
- *die habn verninftiges Geld **for** ihr Ernte gegriegt*
  'sie bekamen gutes Geld **für** ihre Ernte'
- *da war immer ne Masse **vor** die Kinder zu tun*
  'Es gab immer viel zu tun **für** die Kinder'

Verwendung von Abkürzungen und/oder Zeichensetzung
- ***St.** John's* vs. ***Saint** Johns*

QUELLEN

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

Boas, Hans C. 2016. Variation im Texasdeutschen: Implikationen für eine vergleichende Sprachinselforschung. In Alexandra Lenz (ed.), *German Abroad. Perspektiven der Variationslinguistik, Sprachkontakt- und Mehrsprachigkeitsforschung*, 11-44. Vienna 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.

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>

Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10, 353-402.

METHODE

Auf der Basis der Daten des Texas-German-Dialect-Projects (TGDP, Boas et al. 2010) wird ein einheitliches Annotationssystem entwickelt.

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 System basiert auf der Auswertung eines Korpus bestehend aus TGDP **Gesprächsdaten**. Dazu wurden die Datensätze von **150 randomisiert ausgewählten SprecherInnen** auf verschiedenen Ebenen annotiert.

ANFORDERUNGEN & VORAUSSETZUNGEN FÜR DIE ANNOTATION

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

TOOLS

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

DAS SYSTEM IN DER PRAXIS

ZIELE

Hauptziel: Vergleichbarkeit der Daten. Zu diesem Zweck gilt es, eine Annotationsebene zu etablieren, auf der Variation systematisch reduziert wird (z.B. Schreibvarianten eines „Wortes"). Es ist *nicht* das Ziel, die Texasdeutschen Texte zu ‚korrigieren'.

ANNOTATIONEN: MULTI-EBENEN-STRUKTUR

Sprach-Tagging (lang)
- basiert auf Phonologie, Morphologie & Semantik
- Hauptkategorien (Tags): deu (Deutsch), eng (Englisch), deu.txg (deutsches Material mit nicht-standard Bedeutung), mix (z.B. *gejumpt*), NE (Named-Entity), ambig (*), unvollständig (#), unklar (xxx)

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

	341 [01:4	342 [01:4	343 [01:42	344 [01:42.3*]	345 [01:4	346 [01:4	347 [01:4	348 [01:4	349 [01:4	350 [01:4	351 [01:4	352 [01:4	353 [01:4	354 [01:44	355 [01:4
S0035 [tok]	**die**	**erste**	**graden**		**von**	**die**	**schule**	**warn**	**sie**	**auch**	**uh**	**in**	**in**	**deutsch**	**gelern**
S0035 [lang]	deu	deu	*deu.txg	*mix:deu+eng	deu	deu	deu	deu	deu	deu	xxx.hes	deu.txg	deu.txg	deu	deu.txg
S0035 [norm]	die	erste	*Grade	*graden	von	die	Schule	waren	sie	auch	uh	in	in	Deutsch	gelernt
S0035 [trans_deu_utt]	In den unteren Klassenstufen der Schule wurde auch, äh, auf, auf Deutsch unterrichtet.														

Abb. 1. Screenshot des EXMARaLDA Partitur-Editors (Datei: TGDP 1-35-1-20-a)

Es ist ambig, ob es sich bei diesem Token um eine deutsche Form mit einer nicht standardisierten Bedeutung und einer nicht standardisierten Pluralform handelt (→ deu.txg) oder ob der Sprecher eine wortinterne Mischung (englisches 'grade' + deutsche Pluralmarkierung 'n') verwendet (→ mix:deu+eng).

Das Sternchen (*) zeigt an, dass dieses Token auf der tok-Ebene mehrere potenzielle Sprach-Tags und/oder Normalisierungsformen hat.

Grammatikalische Abweichungen, die sich nicht auf die Wortart bzw. das Lemma eines Tokens auswirken, werden in der Norm-Ebene nicht verändert

Verzögerungslaute haben eine Markierung "unklare Sprache" (xxx) mit der Unterkategorie "hes", die anzeigt, dass sie zur Kategorie "hesitation" (=Zögern) gehören.

Im Allgemeinen werden Token als "deu.txg" kategorisiert, wenn es sich um eine deutsche Form mit einer Nicht-Standard-Bedeutung handelt (und ggf. durch ein anderes, standardnäheres Token ersetzt werden könnte). Hier: *in* könnte durch *auf* ersetzt werden und *gelernt* könnte durch *gelehrt* / *unterrichtet* ersetzt werden.

VORTEILE DES SYSTEMS

Sprach-Tagging (lang)
- Wortbasierte Sprachkontaktphänomene sind auffindbar, ohne sie von vornherein zu kategorisieren
- Sprachzuordnung ist für BenutzerInnen nachvollziehbar
- Flexibilität ermöglicht, mehrere Sprachen bzw. Dialekte zu annotieren

Normalisierung (norm)
- zu komplexe 'Korrekturen'/ Interpretationen werden vermieden (→ einheitlicher / konsistenter)
- Das System erlaubt es, unterschiedliche Interpretationen ambiger Tokens zu dokumentieren

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

FAZIT

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

Das vorgestelle System ist ein erster Schritt zur Etablierung eines Standards für die systematische Annotation, das es vergleichbare Abfragen ermöglicht und Annahmen und Interpretationen hinter Normalisierungsentscheidungen transparent und falsifizierbar macht.