Variationslinguistische Analyse mit minimalem händischen Annotationsaufwand

Jan Gorisch
Leibniz-Institut für Deutsche Sprache, Mannheim

Traditionell wird in der Variationslinguistik relativ großer Aufwand betrieben, um Sprachkorpora händisch mit detaillierten Annotationen zu versehen. Das Ziel sind oft Karten auf denen vermerkt ist wo wie gesprochen wird, wie z.B. im „Atlas zur Aussprache des deutschen Gebrauchsstandards“ (AADG, [1]), dessen Ergebnisse auf dem Korpus „deutsch heute“ (DH) beruhen [2]. Das Präsentieren der Ergebnisse allein, schränkt die Nutzbarkeit der Daten für die linguistische Forschungscommunity allerdings sehr ein. Eine gewisse Abhilfe schaffte die Öffnung der Primärdaten für die Forschungsgemeinschaft über die Datenbank für Gesprochenes Deutsch (DGD, [3]). Allerdings beschränken sich hier die Analysemethoden auf das Browsing (z.B. Audios, Videos, Metadaten), die Recherche auf Wort-Ebene (angereichert durch Normalisierung, Lemmatisierung und POS-Tags) und den Export der Trefferliste. Neuerdings sind zwar weitere Prototypen zur Analyse entstanden, aber sie konzentrieren sich hauptsächlich auf gesprächslinguistischen Themen (vgl. [4,5]). Mit diesem Paper haben wir vorrangig also zwei Ziele: Erstens möchten wir aufzeigen welche Möglichkeiten in Korpora wie „deutsch heute“ stecken und dass die sprachlichen Phänomene noch längst nicht ausgeschöpft sind. Und Zweitens möchten wir zeigen, dass viele Analysen auch ohne manuelle Transkription auf Phon-Ebene möglich sind. Wir stellen also ein proof of concept vor, das die reichhaltige Landschaft an Werkzeugen verwendet, wie sie über die letzten Jahrzehnte in der linguistischen – und im Speziellen der phonetischen – Forschung entwickelt wurden.

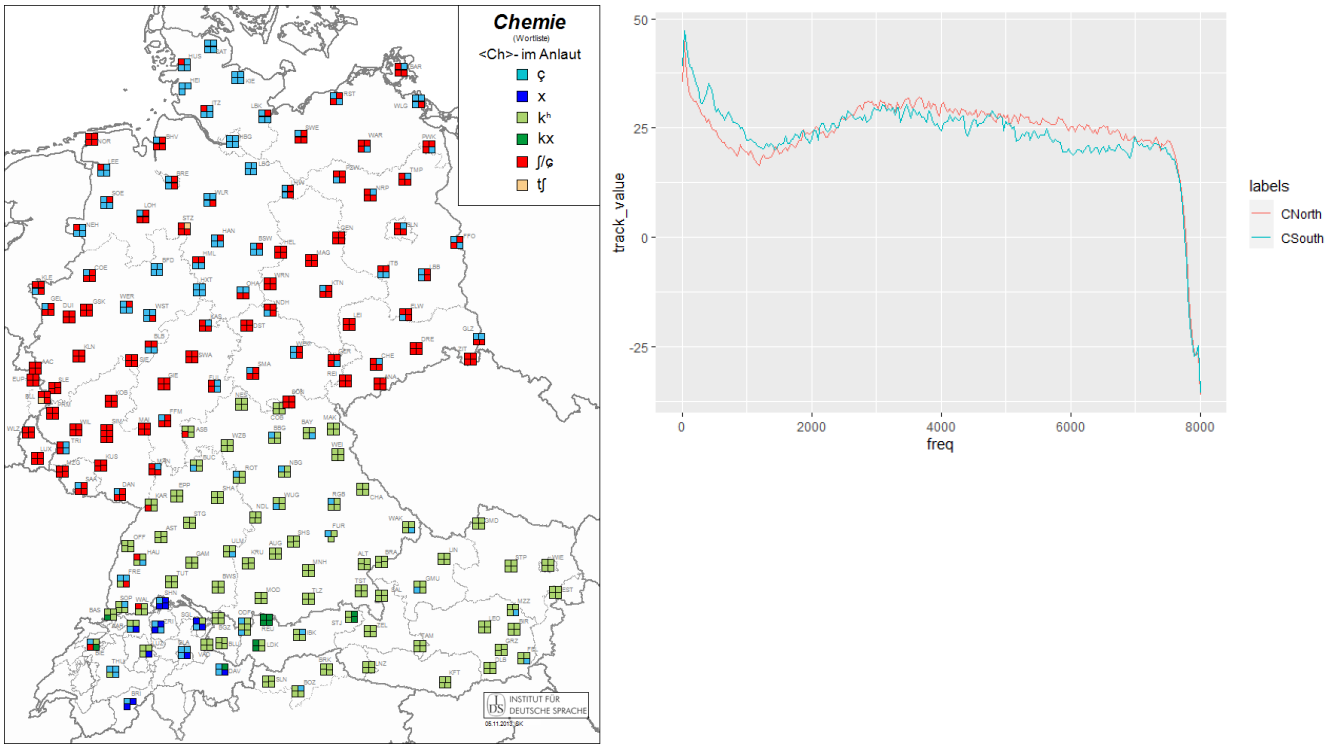
Wir beschränken uns hier auf das Ziel-Phänomen <ch> im Anlaut von Wörtern wie „Chemie“, das vielfältig realisiert werden kann, aber überwiegend als palataler Frikativ [ç], velarer/uvularer Frikativ [x]/[χ], oder als velarer Plosiv [k^h] (häufig auch als palatale Affrikate [kç] realisiert). Wir nehmen an, dass sich diese Sprachlaute spektral in ihrem center of gravity unterscheiden, und dass die Realisierungen regional abhängig sind: mehr palatale Frikative im Norden des deutschsprachigen Raumes vs. mehr velare/uvulare Frikative oder Plosive im Süden. Dies geht aus der entsprechenden Karte (s. Figure 1, links) hervor. Unser Ziel ist, dieses Ergebnis der Kartierung annähernd durch die Anwendung von EMU-R [6] und die darin integrierten Methoden der Signalverarbeitung zu reproduzieren. (Um einen Auszug der reichhaltigen Metadaten des DH-Korpus in EMU-R verwenden zu können, verwendeten wir eine Implementierung der Metadaten-Funktionalität von Fredrik Karlsson [7], welche zum Zeitpunkt dieser Arbeit noch nicht in das offizielle Release von EMU-R eingegangen war.)

Ein bislang wenig beachtetes Sub-Korpus von DH sind die Aufnahmen zur Bildbenennung, welche wir für diese Zwecke erstmals vorverarbeiteten (136 Transkripte auf orthographischer Ebene korrigiert), in eine EMU-Datenbank importierten, und darin die Anbindung zu BAS-webServices [8] nutzten (G2P und MAUS). Als Ziel-Phonem wählten wir das erste im Wort „Chemikerin“ mit 91 Treffer. Weitere Schritte innerhalb EMU-R orientierten sich an Kapitel 21 des Handbuches [9], nämlich einem Rezepts zur spektralen Analyse. Im Speziellen vergaben wir die Labels „CSouth“ und „CNorth“ („C“ für die kanonische Aussprache des palatalen Frikativs in SAMPA), anhand der Breitengrad-Informationen der DH-Geo-Metadaten: alle Ereignisse oberhalb von 49,3° zählten wir zum Norden. Die Ergebnisse sind in Figure 1 (rechts) dargestellt. Es lassen sich, trotz der hart gezogenen Trennlinie zwischen Nord und Süd, die Tendenz erkennen, dass bei den Anfangs-Lauten im Süden, die spektralen Anteile im oberen Bereich, zwischen 4 und 8 kHz gegenüber dem Norden, reduziert sind.

In Zukunft wollen wir genauere Geo-Analysen durchführen, z.B. Karten per Kriging erstellen, und die vorhandenen Variations-Korpora über ein passendes GUI der Forschungsgemeinschaft und ihren speziellen Fragestellungen zugänglich machen.

doi:10.11576/pundp2022-1013

Figure 1. *Karte zum <Ch>-Anlaut in „Chemie“ aus dem AADG (2013), links. Summierte spektrale Schnitte für <Ch>-Anlaut in „Chemikerin“ aus den Bildbenennungen des Deutsch-heute Korpus, rechts.*



[1] Kleiner, S., Berend, N., Brinckmann, C., & Knöbl, R. (2011): "Deutsch heute". Ein sprachgebietsweites Forschungsprojekt zur regionalen Variation in der gesprochenen deutschen Standardsprache. In: *Pohl, Heinz-Dieter (Hrsg.): Akten der 10. Arbeitstagung für bayerisch-österreichische Dialektologie in Klagenfurt, 2007. (= Klagenfurter Beiträge zur Sprachwissenschaft, Jg. 34-36).* Wien: Praesens, 2011. S. 179-193.

[2] Projektergebnisse im Internet (seit 2011) auf der Wiki-Plattform "Atlas zur Aussprache des deutschen Gebrauchsstandards (AADG)": *http://prowiki.ids-mannheim.de/bin/view/AADG/*

[3] Schmidt, T. (2014): The Database for Spoken German – DGD2. In: *Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland: European Language Resources Association (ELRA).

[4] https://zumult.org/

[5] *http://zumult.ids-mannheim.de/ProtoZumult/index.jsp*

[6] Winkelmann, R., Jänsch, K., Cassidy, S. & Harrington, J. (2021). *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.3.0.

[7] https://github.com/samgregory/emuR

[8] Kisler, T., Reichel, U. D., & Schiel, F. (2017): Multilingual processing of speech via web services, Computer Speech & Language, Volume 45, September 2017, pages 326–347.

[9] https://ips-lmu.github.io/The-EMU-SDMS-Manual/recipe-spectralAnalysis.html

doi:10.11576/pundp2022-1013