Das RIDGES-Korpus. Ein diachrones, tief annotiertes Mehrebenenkorpus aus Kräuterkundetexten

Anke Lüdeling, Carolin Odebrecht, Uwe Springmann

Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin

{anke.luedeling, carolin.odebrecht, uwe.springmann} (at) hu-berlin.de

1. Die Forschungsfrage

Ziel des RIDGES-Projekts (**R**egister **i**n **D**iachronic **Ge**rman **S**cience) ist die **qualitative und quantitative Analyse der Entstehung eines deutschsprachigen wissenschaftlichen Registers** [3]. Es handelt sich dabei um derzeit 24 Kräutertexte, die zwischen 1487 und 1870 veröffentlicht wurden und von denen jeweils etwa 30 Seiten in mehreren Ebenen annotiert wurden. Solche kräuterkundlichen Texte sind über den ganzen Zeitraum verfügbar und behandeln ein Gebiet, das anders als z.B. Astrologie/Astronomie einem geringeren Wandel der wissenschaftlichen Methode unterliegt.

Da sich der Übergang von Latein als Medium des wissenschaftlichen Diskurses zu Deutsch über 300 Jahre hinzog [4], ist ein diachrones Korpus für eine empiriegestützte Beantwortung dieser Frage notwendig. In einem solchen Zeitraum wird die Entstehung eines neuen Registers auch durch Änderungen des Lexikons, der Orthographie, der Wortbildung, der Syntax und Textstruktur beeinflusst, die sich mit dem jeweiligen Stand der Drucktechnik und Typographie in unterschiedlicher Weise graphisch manifestierten (Abb. 1, Abb. 2). Zudem vollzogen sich diese Änderungen vor dem Hintergrund eines sich wandelnden Schul- und Universitätssystems sowie der wissenschaftlichen Methode.

Daraus ergeben sich abgeleitete Forschungsfragen zur Konstruktion eines Korpus:

- Welche Art der Transkription und Normalisierung sind für ein solches Korpus notwendig?
- Welche Annotationskategorien (z.B. Wortarten) sind über die Zeit konstant und bezeichnen dieselben Konzepte?
- Was für eine Korpusarchitektur ist notwendig?
- Wie kann die Vergleichbarkeit zu anderen (historischen und modernen) Korpora hergestellt werden?
- Wie kann ein solches Korpus für andere Fragestellungen wiederverwendbar gemacht werden?

2. Das RIDGES-Korpus

RIDGES Version 4.1:

- 29 Exzerpte aus 24 Texten (1487-1870)
- pro Exzerpt etwa 5000 Wörter (insgesamt 154.267 Wörter)
- frei verfügbar unter `http://korpling.german.hu-berlin.de/ridges/index_de.html` und im LAUDATIO-Repository: `http://www.laudatio-repository.org/repository/`
- wurde von Studenten erstellt
- wird ständig erweitert (Version 4.2: 6 neue Texte)
- gefördert durch zwei Google Digital Humanities Awards

3. Korpusdesign und -architektur

Notwendigkeit einer Annotation auf mehreren Ebenen mit multipler Segmentierung [1] (Abb. 1, Abb. 3):

- diplomatische Ebene (vnd, vñ, d̊, vonn, menſes)
- clean-Ebene (vnd, vnd, der, vonn, menses + Worttrennungen aufgehoben)
- normalisierte Ebene (und, und, der, von, menses)
- Wortart-Annotierung auf Basis der Norm-Ebene (STTS-Tagset [5], TreeTagger [6]) (Tab. 1)
- Syntax-Annotierung auf Basis von Token-Spannen (Tab. 1)
- Annotationen und Normalisierungen sind voneinander unabhängig (wichtige Voraussetzung für Wiederverwendung)

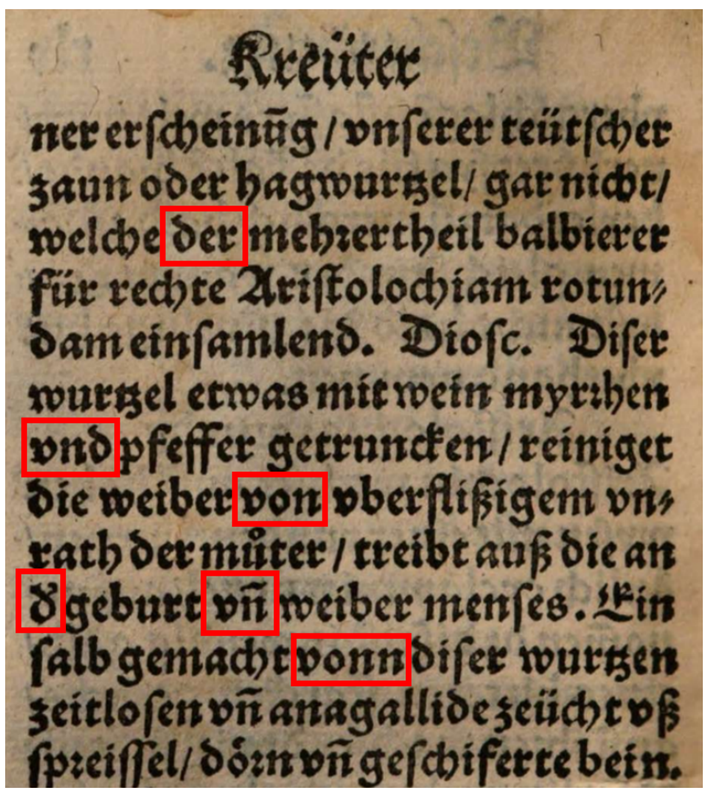


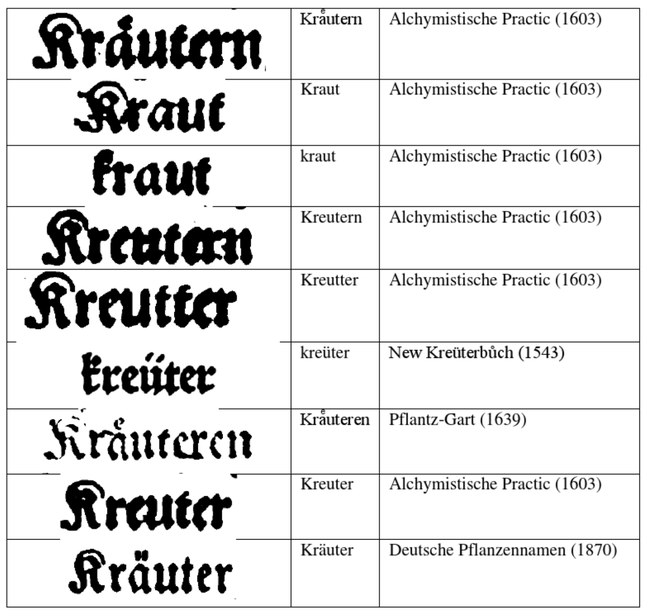
Abbildung 1: *Typographisch (Randausgleich!) motivierte Variation der Wörter und, der, von (Bodenstein, Wie sich meniglich, 1557). Griechisch-lateinische Pflanzennamen wie Anagallis werden im Satzzusammenhang in deklinierter Form verwendet (anagallide).*

Diplomatische Transkription unter Verwendung von UTF-8 Codesequenzen der letzten Zeilen aus Abb. 1:

Ein
ſalb gemacht vonn diſer wurtzen
zeitloſen vñ anagallide zeücht vß
ſpreiſſel/ dörn vñ geſchiferte bein.

Normalisiert:

Eine
Salbe gemacht von dieser Wurzel
Zeitlosen und Anagallis zieht aus
Spreißel/ Dornen und geschieferte Bein.
(geschieferte Bein = gesplitterte Knochen)



Kräutern	Kräutern	Alchymistische Practic (1603)
Kraut	Kraut	Alchymistische Practic (1603)
kraut	kraut	Alchymistische Practic (1603)
Kreutern	Kreutern	Alchymistische Practic (1603)
Kreutter	Kreutter	Alchymistische Practic (1603)
kreüter	kreüter	New Kreüterbůch (1543)
Kräuteren	Kräuteren	Pflantz-Gart (1639)
Kreuter	Kreuter	Alchymistische Practic (1603)
Kräuter	Kräuter	Deutsche Pflanzennamen (1870)

Abbildung 2: *Schreibvariation von Kraut/Kräuter im Korpus [3].*

dipl	/	reiniget	die	weiber	von	vberfliſzigem	vnꝛ	rath	der	mûter	/
clean	/	reiniget	die	weiber	von	vberfliszigem	vnrath		der	muter	/
norm	/	reinigt	die	Weiber	von	überflüssigem	Unrat		der	Mutter	/

Abbildung 3: *Drei Transkriptionsebenen: diplomatisch, clean, normiert (Bodenstein 1557). Visualisierung mit der Korpus-Visualisierungs-Software ANNIS [2]: http://www.corpus-tools.org. Online-Link: `https://korpling.german.hu-berlin.de/annis3/?id=7d224ec2-562c-4797-bc39-ab2e73f4fe31`*

Tabelle 1: *Beispielannotation eines Satzes aus Tallat, Artzney Buchlein der Kreutter, 1532*

Token	*den*	*Samen*	*trinck*	*mit*	*Venchel*	*Waſſer*
POS	ART	NN	VVFIN	APPR	NN	
Syntax	NP			PP		

4. Erste Ergebnisse

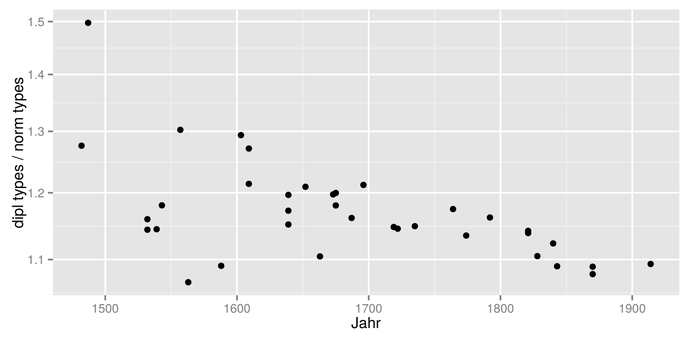


Abbildung 4: *Mittlere Anzahl verschiedener Schreibweisen pro normierter Wortform für die einzelnen Werke in RIDGES. Bodenstein (1557) liegt bei 1,3 (Abb. 1). Im Laufe der Zeit nimmt die Varianz innerhalb der Werke deutlich ab.*

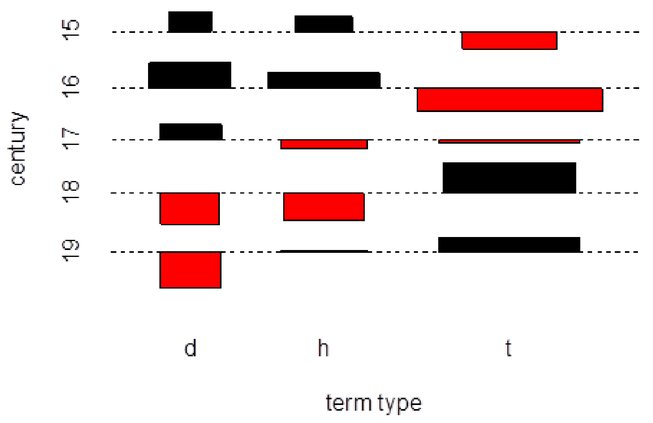


Abbildung 5: *Assoziationsplot für die Häufigkeit von Fachbegriffen: Kräuter (h: herbs), Krankheiten (d: diseases), andere Fachbegriffe (t: termini technici) im Zeitverlauf. Bei der gestrichelten Grundlinie entsprechen die beobachteten Häufigkeiten der Annahme der statistischen Unabhängigkeit der Terme vom Zeitraum. Oberhalb der Linie sind Terme überhäufig, darunter unterhäufig. Je größer der vertikale Abstand von der Grundlinie, desto unwahrscheinlicher ist die angenommene Unabhängigkeit. In den früheren Jahrhunderten werden hauptsächlich Pflanzen und Krankheiten erwähnt, später hauptsächlich andere Termini (z.B. Baum, Gewächse, Empfinden, Nerven).*

Literatur

[1] Thomas Krause, Anke Lüdeling, Carolin Odebrecht und Amir Zeldes. „Multiple tokenizations in a Diachronic Corpus“. In: *Exploring Ancient Languages through Corpora Conference (EALC). 14.-16.06.2012, Oslo*. 2012.

[2] Thomas Krause und Amir Zeldes. „ANNIS3: A new architecture for generic corpus query and visualization“. In: *Digital Scholarship in the Humanities* (2014), fqu057.

[3] Carolin Odebrecht, Malte Belz, Vivian Voigt, Amir Zeldes und Anke Lüdeling. „RIDGES Herbology - Designing a Diachronic multi-layer Corpus“. In: (Eingereicht).

[4] Uwe Pörksen. „Deutsche Sprachgeschichte und die Entwicklung der Naturwissenschaften.–Aspekte einer Geschichte der Naturwissenschaftssprache und ihrer Wechselbeziehung zur Gemeinsprache“. In: *Besch, W. et al. (Hgg.) Sprachgeschichte. Ein Handbuch der deutschen Sprache und ihrer Erforschung* 1 (1984), S. 85–101.

[5] Anne Schiller, Christine Thielen, Simone Teufel und Christine Stöckert. *Guidelines fürr das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. 1999. URL: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf`.

[6] Helmut Schmid. „Probabilistic part-of-speech tagging using decision trees“. In: *Proceedings of the international conference on new methods in language processing*. Bd. 12. Citeseer. 1994, S. 44–49.

Konferenz: Entstehung und Frühgeschichte der modernen deutschen Wissenschaftssprachen: Vernakuläre Gelehrtenkommunikation in der Frühen Neuzeit, Greifswald, Alfried Krupp Wissenschaftskolleg, 12. bis 14. November 2015