Uwe Springmann, Anke Lüdeling, Carolin Odebrecht
und Thomas Krause

Das RIDGES-Korpus

Ein diachrones, tief annotiertes Mehrebenenkorpus aus Kräuterkundetexten

1 Die Forschungsfrage

Die Forschungsfrage des RIDGES-Projekts (**R**egister **i**n **D**iachronic **Ge**rman **S**cience) lautet: *Wie kann man die Entstehung eines deutschsprachigen wissenschaftlichen Registers qualitativ und quantitiv analysieren?* (Odebrecht et al. 2017). Es handelt sich dabei (Stand Anfang 2017, Version RIDGES 5.0) um 33 Kräuterkundetexte, die zwischen 1482 und 1914 veröffentlicht wurden und von denen jeweils etwa 30 Seiten in mehreren Ebenen annotiert wurden. Solche kräuterkundlichen Texte sind über den gesamten Zeitraum verfügbar und behandeln ein Gebiet, das anders als z. B. Astrologie/Astronomie einem geringeren Wandel der wissenschaftlichen Methode unterlag (Gloning 2007; Habermann 2001).

Da sich der Übergang von Latein als Medium des wissenschaftlichen Diskurses zu Deutsch über einen Zeitraum von 300 Jahren hinzog (Pörksen 1984), ist ein *diachrones* Korpus für eine empiriegestützte Beantwortung dieser Frage notwendig.

In einem solchen Zeitraum wird die Entstehung eines neuen Registers, das wir im Sinne von (Biber & Conrad 2009) und anderen als multidimensional verstehen und das daher auf allen Ebenen (gleichzeitig) untersuchbar sein

Danksagung: Wir danken unseren Kollegen Laura Perlitz, Gohar Schnelle, Malte Belz, Felix Golcher sowie unseren vielen Studierenden für ihre tatkräftige Unterstützung beim Aufbau des RIDGES-Korpus!

Uwe Springmann, Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität München; Institut für deutsche Sprache und Lingustik, Humboldt-Universität zu Berlin, E-Mail: uwe@springmann.net
Anke Lüdeling, Institut für deutsche Sprache und Lingustik, Humboldt-Universität zu Berlin, E-Mail: anke.luedeling@rz.hu-berlin.de
Carolin Odebrecht, Institut für deutsche Sprache und Lingustik, Humboldt-Universität zu Berlin, E-Mail: carolin.odebrecht@hu-berlin.de
Thomas Krause, Institut für deutsche Sprache und Lingustik, Humboldt-Universität zu Berlin, E-Mail: krauseto@rz.hu-berlin.de

https://doi.org/10.1515/9783110476958-020

muss, auch durch Änderungen des Lexikons, der Orthographie, der Wortbildung, der Syntax, der Textstruktur etc. beeinflusst, die sich mit dem jeweiligen Stand der Drucktechnik und Typographie in unterschiedlicher Weise graphisch manifestierten (Abb. 2, Abb. 3).

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

Zur Beantwortung der oben gestellten Fragen wird seit einigen Jahren das RIDGES-Korpus am Lehrstuhl für Korpuslinguistik und Morphologie (Anke Lüdeling) an der Humboldt-Universität zu Berlin im Rahmen von Veranstaltungen zur historischen Korpuslinguistik von Studierenden erstellt. Die 33 Texte von RIDGES 5.0 stellen Extrakte von je zwischen 3.000 und 10.000 Wörtern aus umfangreichen Kräuterkundebüchern dar, die zwischen 1482 und 1914 erschienen sind und insgesamt einen Umfang von 180.000 Wörtern haben. Sie wurden nach ausführlichen Richtlinien[1] transkribiert und in einer Vielzahl von Ebenen annotiert.[2] Da die betreffende Lehrveranstaltung eine ständige Einrichtung ist, die einmal pro Jahr stattfindet, wird das Korpus laufend erweitert. Es ist unter der offenen Lizenz CC-BY frei verfügbar.[3]

1 https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/ridgesv5_2016-10-19.pdf [letzter Zugriff: 30. 09. 2018].

2 Diese Arbeit wurde zweimal durch einen Google Digital Humanities Award gefördert.

3 Jede neue Version des Korpus ist im LAUDATIO-Repository langfrist archiviert und referenziert. Für Version 5.0: Lüdeling, Anke; Odebrecht, Carolin; Zeldes, Amir; RIDGES-Herbology (Version 5.0), Humboldt-Universität zu Berlin. http://korpling.german.hu-berlin.de/ridges/. http://hdl.handle.net/11022/0000-0001-C98F-C [letzter Zugriff: 30. 09. 2018].

Zusätzlich findet eine Erweiterung des RIDGES-Korpus durch die Verwendung moderner, auf neuronalen Netzen aufbauender OCR-Methoden statt, bei der eine Texterkennung auf kompletten RIDGES-Titeln durchgeführt wird (Springmann & Lüdeling 2016). Dieses Erweiterungskorpus steht unter dem Namen RIDGES-OCR zur Verfügung.[4]

3 Korpusdesign und -architektur

Aufgrund der Komplexität des Druckmaterials ist es sehr vorteilhaft, die Annotierung auf mehreren voneinander unabhängigen Ebenen vorzunehmen. Dadurch ist es z. B. möglich, multiple Segmentierungen vorzunehmen (Krause et al. 2012). So gibt es in RIDGES u. a. folgende Ebenen (siehe Abb. 2, Abb. 1):

- diplomatische Ebene (Beispiel: *vnd, vñ, d̊, vonn, menſes*)
- clean-Ebene (*vnd, vnd, der, vonn, menses*; Worttrennungen sind aufgehoben)
- normalisierte Ebene (*und, und, der, von, menses*)
- Wortart-Annotierung auf Basis der Norm-Ebene (STTS-Tagset: Schiller et al. 1999, TreeTagger: Schmid 1994; siehe Abb. 1)
- weitere Annotierung auf Basis von Token-Spannen (Abb. 1)

dipl	in	Clyftieren	hats		nicht	ein	geringe	Ehr	.
clean	in	Clystieren	hats		nicht	ein	geringe	Ehr	.
norm	in	Klistieren	hat	es	nicht	eine	geringe	Ehre	.
lemma	in	<unknown>	haben	es	nicht	eine	gering	Ehre	.
pos	APPR	NN	VAFIN	PPER	PTKNEG	ART	ADJA	NN	$.
erlaeuterung		Einläufen							
figure	figure								
lb	lb								
pb	pb								
pb_n	310								
typeface	gothic	gothic	gothic		gothic	gothic	gothic	gothic	gothic

Abb. 1: Transkriptionsebenen (diplomatisch, clean, normiert), Lemma, Wortart (pos) und Spannenannotation (Zeile: lb; Seite: pb; Seitenzahl: pb_n; Schriftart: Fraktur) aus Bodenstein 1557. Visualisierung mit der Korpus-Visualisierungs-Software ANNIS (Krause & Zeldes 2016): Siehe http://www.corpus-tools.org und für diese Grafik: https://korpling.org/annis3/?id=23dde714-37bc-4576-9923-c7d0b71f65d8.

4 RIDGES-OCR: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/ocr [letzter Zugriff: 30. 09. 2018].

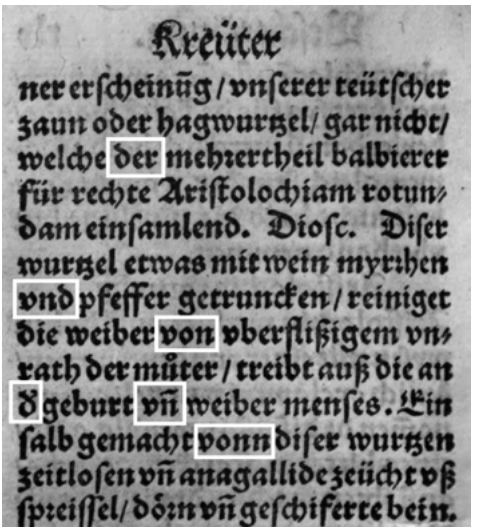

Abb. 2: Typographisch (Randausgleich!) motivierte Variation der Wörter *und*, *der*, *von* (Bodenstein, Wie sich meniglich, 1557). Griechisch-lateinische Pflanzennamen wie Anagallis werden im Satzzusammenhang in deklinierter Form verwendet (anagallide).

Annotationen und Normalisierungen sind dabei voneinander unabhängig, was eine wichtige Voraussetzung für die Wiederverwendung des Korpus ist.

Mit Hilfe der unterschiedlichen Annotationsebenen kann man sich den Text sowohl in diplomatischer Transkription unter Verwendung von UTF-8 Codesequenzen als auch in moderner normalisierter Schreibweise ausgeben lassen. Als Beispiel seien hier die letzten Zeilen aus Abb. 2 angegeben:

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

In normalisierter moderner Schreibweise stellt sich das wie folgt dar:

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

Diese Normalisierung ermöglicht auch eine gezielte Suche nach Wörtern, die in vielen verschiedenen historischen Schreibweisen im Korpus vorhanden sind. Dabei treten nicht nur chronologische oder dialektale Varianten auf (Abb. 3), sondern es finden sich auch mehrere Varianten eines Wortes im selben Werk, ja sogar auf derselben Seite (Abb. 2) aus augenscheinlich rein typographischen Beweggründen.

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)

Abb. 3: Schreibvariation von Kraut/Kräuter im Korpus (Odebrecht et al. 2017).

4 Erste Ergebnisse

Als eines der ersten Ergebnisse präsentieren wir (Abb. 4) die Varianz der Schreibweise im Zeitverlauf. Abgebildet ist die mittlere Anzahl der Schreibvarianten, die zu einer gemeinsamen normalisierten Wortform gehören. Jeder Punkt entspricht einem Text, der über sein Erscheinungsdatum einer Zeit zuge-

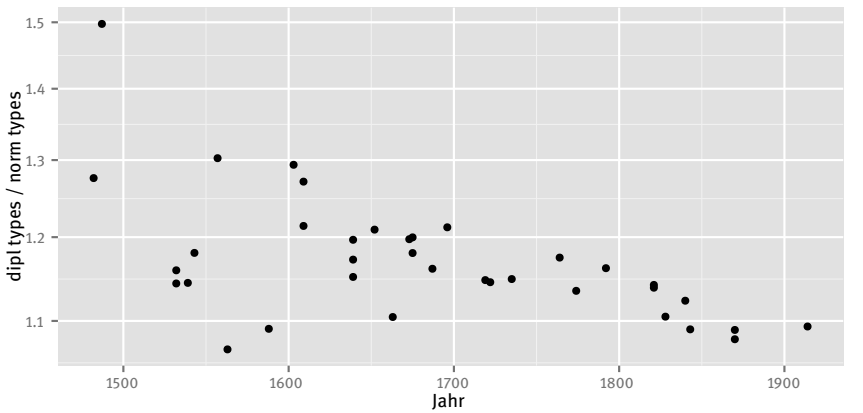

Abb. 4: Mittlere Anzahl verschiedener Schreibweisen pro normierter Wortform für die einzelnen Werke in RIDGES. Bodenstein (1557) liegt bei 1,3 (Abb. 2). Im Laufe der Zeit nimmt die Varianz innerhalb der Werke deutlich ab (Odebrecht et al. 2017).

ordnet werden kann. Es ist klar zu erkennen, dass die Schreibvariation im Lauf der Zeit abnimmt und sich einer normierten Schreibung annähert.

Viele weitere Ergebnisse zu unterschiedlichen linguistischen Ebenen finden sich in Odebrecht et al. (2017) und in Abschlussarbeiten (u. a. Perlitz 2014).

Reference

Biber, Douglas & Susan Conrad (2009): *Register, Genre, and Style*. Cambridge University Press.

Gloning, Thomas (2007): Deutsche Kräuterbücher des 12. bis 18. Jahrhunderts. Textorganisation, Wortgebrauch, funktionale Syntax. In: Andreas Meyer und Jürgen Schulz-Grobert (Hrsg.): *Gesund und krank im Mittelalter. Marburger Beiträge zur Kulturgeschichte der Medizin*. Leipzig: Eudora-Verlag, 9–88.

Habermann, Mechthild (2001): Frühneuzeitliche Kräuterbücher und ihre makrostrukturelle Textorganisation. In: *Deutsche Fachtexte der frühen Neuzeit: naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache*. Berlin: Walter de Gruyter, 98–167.

Krause, Thomas & Amir Zeldes (2016): Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1), 118–139. URL: http://dsh.oxfordjournals.org/content/31/1/118 [letzter Zugriff: 30. 09. 2018].

Krause, Thomas, Anke Lüdeling, Carolin Odebrecht & Amir Zeldes (2012): Multiple tokenizations in a diachronic corpus. In: *Exploring Ancient Languages through Corpora Conference (EALC)*.

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling & Thomas Krause (2017): Ridges herbology – designing a diachronic multi-layer corpus. *Language Resources and Evolution* 51(3), 695–725. DOI: 10.1007/s10579-016-9374-3

Perlitz, Laura (2014): *Konkurrenz zwischen Wortbildung und Syntax – historische Entwicklung von Benennung*. Philosophische Fakultät II der Humboldt-Universität zu Berlin. URL: http://edoc.hu-berlin.de/docviews/abstract.php?id=41626 [letzter Zugriff: 30. 09. 2018].

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

Schiller, Anne, Christine Thielen, Simone Teufel & Christine Stöckert (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). URL: http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf [letzter Zugriff: 30. 09. 2018].

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees (1994). In: *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Springmann, Uwe & Anke Lüdeling (2016): OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly* 11(2). URL: http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.