Das ZDL-Regionalkorpus

Ein Korpus für die lexikografische Beschreibung der diatopischen Variation im Standarddeutschen

Andreas Nolda/Adrien Barbaresi/Alexander Geyken

Das ZDL-Regionalkorpus umfasst Zeitungsartikel aus Lokal- und Regionalressorts deutschsprachiger Tageszeitungen. Es dient als empirische Grundlage für die lexikografische Beschreibung der diatopischen Variation im *Digitalen Wörterbuch der deutschen Sprache* (DWDS). Darüber hinaus steht es allen angemeldeten Nutzern der DWDS-Korpusplattform für die Recherche zur Verfügung. Die Abfrage kann auf bestimmte diatopische Areale oder diachrone Zeiträume beschränkt werden. Die Verteilung der Treffer über Areale und Zeiträume lässt sich in verschiedener Form darstellen; dabei werden neben absoluten Trefferzahlen auch normalisierte PPM-Werte ausgegeben.

1 Korpusdesign

Das *ZDL-Regionalkorpus* des Zentrums für digitale Lexikographie der deutschen Sprache ist ein Korpus von Zeitungsartikeln aus Lokal- und Regionalressorts deutschsprachiger Tageszeitungen. Artikel anderer Ressorts wurden ausgeschlossen, da diese oft von Nachrichtenagenturen oder überregionalen Zentralredaktionen stammen. Das Korpus dient den Lexikografen des *Digitalen Wörterbuchs der deutschen Sprache* (DWDS) als empirische Grundlage für die Beschreibung der diatopischen Variation im deutschen Gebrauchsstandard. Auf der Korpusplattform des DWDS (Geyken et al. 2017) können unter www.dwds.de/d/k-meta#regional auch alle angemeldeten Nutzer im ZDL-Regionalkorpus recherchieren.

In den Metadaten sind die Zeitungsartikel je einem Land sowie einem Areal aus der Arealklassifikation des *Variantenwörterbuchs des Deutschen* (Ammon et al. 2016) und der *Variantengrammatik des Standarddeutschen* (2018) zugeordnet. In der aktuellen Korpusversion (Stand: Mai 2020) sind die Areale D-Nordwest, D-Nordost, D-Mittelwest, D-Mittelost, D-Südwest und D-Südost durch je drei bis vier Zeitungen abgedeckt. Der regelmäßig aktualisierte Datenbestand umfasst

gegenwärtig 20,9 Millionen Artikel mit 6,3 Milliarden Tokens aus dem Zeitraum von 1993 bis 2020. Da der im Korpus enthaltene Archivbestand der einzelnen Zeitungen unterschiedlich weit in die Vergangenheit zurückreicht, gibt es erst ab 2005 Daten aus allen Arealen (16,4 Millionen Artikel mit 4,9 Milliarden Tokens) und ab 2017 Daten aus allen Zeitungen (4,1 Millionen Artikel mit 1,3 Milliarden Tokens). Für zukünftige Versionen des Korpus ist eine Erweiterung um Zeitungen aus Österreich und der deutschsprachigen Schweiz vorgesehen.

Die DWDS-Korpusplattform stellt für die Recherche im ZDL-Regionalkorpus komfortable Werkzeuge zur Verfügung. Die Abfrage kann auf bestimmte Areale oder Zeiträume beschränkt werden. Die Verteilung der Treffer über Areale und Zeitungen lässt sich in tabellarischer Form ausgeben. Dabei wird neben der absoluten Trefferzahl jeweils auch ein normalisierter PPM-Wert (*parts per million*) aufgeführt, der die Trefferzahl pro Million Tokens im Areal angibt. Der diachrone Verlauf der PPM-Werte kann in einer parametrisierbaren Histogrammansicht visualisiert werden.

2 Beispieldaten

Zur Illustration seien im Folgenden die Ergebnisse dreier Korpusrecherchen im ZDL-Regionalkorpus wiedergegeben, und zwar nach den Lemmata „schnacken“, „schwätzen“ und „ratschen“ im Zeitraum von 2005 bis 2020 (Stand: Mai 2020). Dies ist, wie in Abschnitt 1 erwähnt, der größte Zeitraum mit Daten aus allen Arealen und deshalb die Default-Einstellung der Histogrammansicht.

Wie Tabelle 1 zeigt, kommt das Lemma „schnacken“ vor allem im Areal D-Nordwest (PPM: 2,79) und – mit etwas niedriger Frequenz – im Areal D-Nordost (PPM: 0,89) vor. Das Lemma „schwätzen“ hingegen ist typisch für die Areale D-Mittelwest (PPM: 2,70) und D-Südwest (PPM: 2,49) (vgl. Tabelle 2). Und das Lemma „ratschen“ kommt fast ausschließlich im Areal D-Südost vor (PPM: 2,06) (Tabelle 3). Eine analoge Verteilung weisen auch die Histogramme in den Abbildungen 1, 2 und 3 auf. Dies gilt zumindest für den Zeitraum ab 2017, für den Daten aus allen Zeitungen zur Verfügung stehen (s. Abschnitt 1).

Die arealen Verteilungen dieser Lemmata im ZDL-Regionalkorpus entsprechen somit denjenigen, die nach den Ergebnissen der siebten Runde des „Atlas zur deutschen Alltagssprache“ (Elspaß/Möller o. J., Frage 8b „über Alltägliches reden“) zu erwarten sind.

Literatur

Ammon, Ulrich et al. (2016): Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland, Liechtenstein, Luxem-

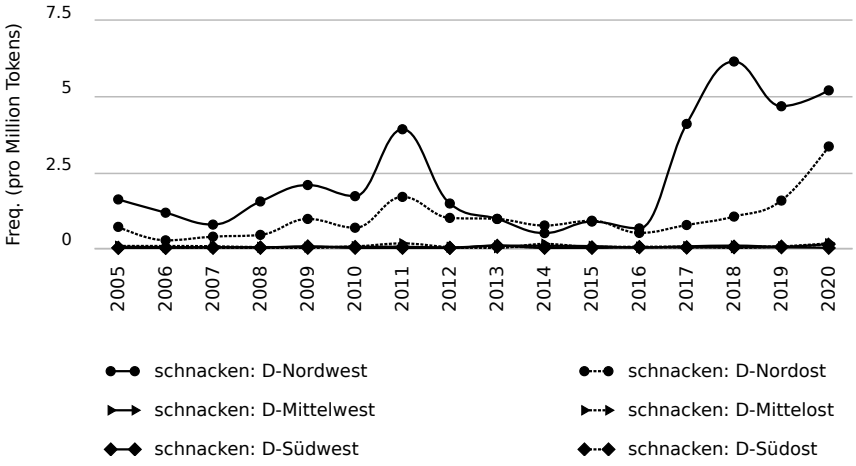

Abbildung 1: diachroner Verlauf der arealen Verteilung von „schnacken“ (Stand: Mai 2020)

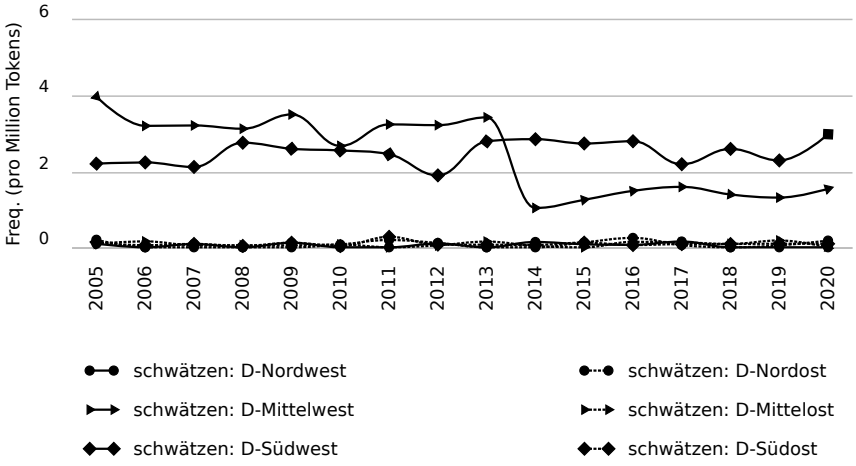

Abbildung 2: diachroner Verlauf der arealen Verteilung von „schwätzen“ (Stand: Mai 2020)

Areal	Treffer	PPM
D-Nordwest	650	2,79
D-Nordost	284	0,89
D-Mittelwest	73	0,04
D-Mittelost	26	0,05
D-Südwest	14	0,02
D-Südost	32	0,03

Tabelle l: Treffer für „schnacken“ im Zeitraum von 2005 bis 2020 (Stand: Mai 2020)

Areal	Treffer	PPM
D-Nordwest	12	0,05
D-Nordost	21	0,07
D-Mittelwest	5164	2,70
D-Mittelost	38	0,08
D-Südwest	2041	2,49
D-Südost	96	0,09

Tabelle 2: Treffer für „schwätzen“ im Zeitraum von 2005 bis 2020 (Stand: Mai 2020)

Areal	Treffer	PPM
D-Nordwest	10	0,04
D-Nordost	11	0,03
D-Mittelwest	89	0,05
D-Mittelost	11	0,02
D-Südwest	53	0,06
D-Südost	2218	2,06

Tabelle 3: Treffer für „ratschen“ im Zeitraum von 2005 bis 2020 (Stand: Mai 2020)

burg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. 2. Aufl. Berlin: Walter de Gruyter.

Elspaß, Stephan/Möller, Robert (o. J.): Atlas zur deutschen Alltagssprache (AdA). http://www.atlas-alltagssprache.de (Stand: 24. Feb. 2020).

Geyken, Alexander et al. (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). In: Zeitschrift für germanistische Linguistik 45, 327–344.

Variantengrammatik des Standarddeutschen: Ein Online-Nachschlagewerk (2018). Verfasst von einem Autorenteam unter der Leitung von Christa Dür-

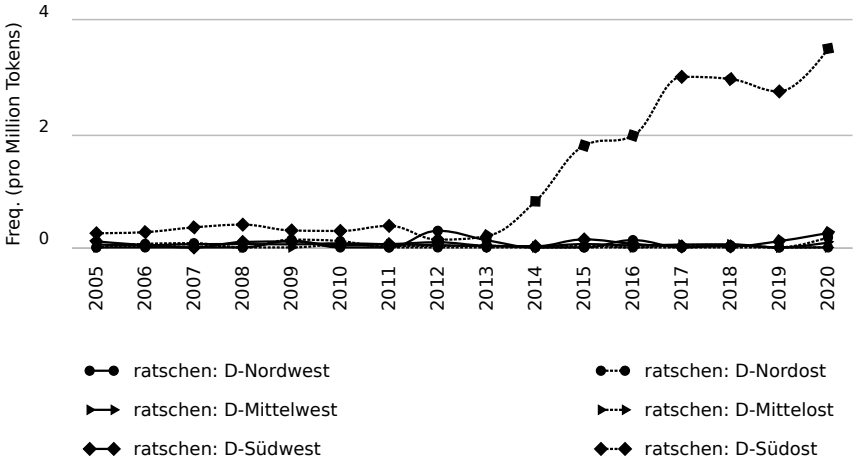

Abbildung 3: diachroner Verlauf der arealen Verteilung von „ratschen“ (Stand: Mai 2020)

scheid, Stephan Elspaß und Arne Ziegler. http://mediawiki.ids-mannheim.de/VarGra/ (Stand: 9. Dez. 2019).

Andreas Nolda
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstraße 22/23
10117 Berlin
andreas.nolda@bbaw.de

Adrien Barbaresi
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstraße 22/23
10117 Berlin
barbaresi@bbaw.de

Alexander Geyken
Berlin-Brandenburgische Akademie der Wissenschaften
Jägerstraße 22/23
10117 Berlin
geyken@bbaw.de