Originalveröffentlichung in: Deppermann, Arnulf/Fandrych, Christian/Kupietz, Marc/Schmidt, Thomas (Hrsg.): Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial. – Berlin/Boston: de Gruyter, 2023. S. 29-52. (Jahrbuch / Leibniz-Institut für Deutsche Sprache 2022)
DOI: *https://doi.org/10.1515/9783111085708*

Andreas Nolda/Adrien Barbaresi/Alexander Geyken (Berlin)

# Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache

## Das ZDL-Regionalkorpus und das Webmonitor-Korpus

**Abstract:** Dieser Beitrag stellt zwei Korpora vor, die als Datengrundlage für die Bestimmung der Regionalangaben im *Digitalen Wörterbuch der deutschen Sprache* (DWDS) fungieren: das *ZDL-Regionalkorpus* und das *Webmonitor-Korpus*. Diese Korpora wurden am Zentrum für digitale Lexikographie der deutschen Sprache (ZDL) erstellt und stehen allen registrierten Nutzern der DWDS-Plattform für Recherchen zur Verfügung. Das ZDL-Regionalkorpus enthält Artikel aus Lokal- und Regionalressorts deutscher Tageszeitungen, die mit arealen Metadaten versehen sind. Es wird ergänzt durch regionale Internet-Quellen im Webmonitor-Korpus, die zusätzliche Areale und Ortspunkte aus dem deutschen Sprachraum einbeziehen. Die Benutzerschnittstelle der linguistisch annotierten Korpora erlaubt nicht nur komplexe sprachliche Abfragen, sondern bietet auch statistische Recherchewerkzeuge zur Bestimmung arealer Verteilungen.

## 1 Überblick

Teil des Arbeitsprogramms des Zentrums für digitale Lexikographie der deutschen Sprache (ZDL) an der Berliner Arbeitsstelle ist die Überarbeitung und Ergänzung der Regionalangaben im *Digitalen Wörterbuch der deutschen Sprache* (DWDS). Insbesondere sollen diese Angaben denjenigen im *Variantenwörterbuch des Deutschen* (Ammon/Bickel/Lenz (Hg.) 2016) angeglichen werden. Da das DWDS korpusbasiert erarbeitet wird und einschlägige Korpora zur lexikographischen Beschreibung diatopischer Variation in der deutschen Standardsprache nicht allgemein verfügbar waren, wurde das *ZDL-Regionalkorpus* erstellt. Dabei wurden Desiderata berücksichtigt, die sich aus dem Design der vergleichbaren, aber nichtöffentlichen Projektkorpora von *Variantenwörterbuch* und *Variantengrammatik* (Variantengrammatik des Standarddeutschen 2018) ergeben. Außerdem wurde ein Arealkonzept erstellt, das die Einteilung von Deutschland in sechs Areale von *Variantenwörterbuch* und *Variantengrammatik* übernimmt und sich hinsichtlich

https://doi.org/10.1515/9783111085708-003

der Grenzverläufe an der korpusbasierten Dialektgliederung von Lameli (2013, S. 194) orientiert. Wie bei diesen Projektkorpora handelt es sich beim ZDL-Regionalkorpus um ein Korpus standardsprachlicher (Zeitungs-)Texte und nicht um eine Sammlung genuin dialektologischer Daten.

Das ZDL-Regionalkorpus enthält gegenwärtig (Mai 2022) 31,5 Mio. Artikel mit insgesamt 9,1 Mrd. Tokens aus Lokal- und Regionalressorts deutscher Tageszeitungen, die mit arealen Metadaten versehen sind. Ergänzt wird das ZDL-Regionalkorpus durch Internet-Quellen mit regionalem Bezug im *Webmonitor-Korpus*, deren Daten nach der Methodik der *Variantengrammatik* (Datenerhebung 2018) erhoben wurden und zusätzliche Areale und Ortspunkte aus dem deutschen Sprachraum einbeziehen. Beide Korpora stehen nicht nur den DWDS-Lexikographen, sondern allen registrierten Nutzern der DWDS-Plattform (Geyken et al. 2017) für eigene Recherchen zur Verfügung. Wie alle Korpora auf der DWDS-Plattform sind die hier beschriebenen Korpora linguistisch annotiert mit Lemmata und Part-of-speech-Tags. Darüber hinaus kann die Verteilung über Areale und Zeitungen abgefragt und im Falle des ZDL-Regionalkorpus auch kartographisch visualisiert werden.

Beide Korpora werden regelmäßig aktualisiert: das ZDL-Regionalkorpus monatlich und das Webmonitor-Korpus täglich. Es handelt sich also technisch gesehen in beiden Fällen um Monitorkorpora. In der Korpuslinguistik wurde die Relevanz von Monitorkorpora früh erkannt (Sinclair 1982). Nach diesem Konzept werden Texte nach und nach verarbeitet und verfügbar gemacht, so dass solche Korpora aktuell gehalten werden. So betrachtet Clear (1987) Monitorkorpora als über die Zeit gleitende Fenster, die immer wieder aktuell gehalten werden, indem die älteren Texte herausgenommen werden. Unser Verständnis ist hingegen, dass solche Korpora allmählich größer werden, indem man alle Datenpunkte behält. Anders als bei früheren Unternehmungen sind mit größeren Korpora einhergehenden technischen Hürden heutzutage nicht mehr so problematisch. Außerdem können diachronische Entwicklungen so festgestellt werden, die u. a. für die Lexikographie von Belang sind.

Dieser Beitrag ist folgendermaßen gegliedert. Abschnitt 2 stellt Desiderata zusammen für Korpora, die als Datengrundlage für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen konzipiert sind. Abschnitt 3 greift diese Desiderata auf und erläutert auf dieser Grundlage Design und Arealkonzept des ZDL-Regionalkorpus. In Abschnitt 4 werden die Benutzerschnittstelle des ZDL-Regionalkorpus und dessen spezifischen Recherchewerkzeuge vorgestellt. In Abschnitt 5 wird in Form einer kleinen Fallstudie überprüft, inwieweit sich in diesem Korpus areale Verteilungen nachweisen lassen, die für den *Atlas zur deutschen Alltagssprache* (Elspaß/Möller 2003–) erhoben wurden. In Abschnitt 6 wird erläutert, wie das ZDL-Regionalkorpus als Datengrundlage für die Beschreibung

diatopischer Markierungen des Digitalen Wörterbuchs der deutschen Sprache (DWDS) verwendet wird. Abschnitt 7 beschreibt die Adaptierung eines Monitorkorpus aus Internetquellen, dessen Erhebung ähnlich wie bei der *Variantengrammatik* verläuft, einige Nutzungsszenarien werden exemplarisch vorgeführt. Der Beitrag schließt in Abschnitt 8 mit einem Ausblick auf die Grenzen von Korpora für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen.

## 2 Desiderata

In einem Aufsatz zur Erstellung der zweiten Auflage des *Variantenwörterbuchs des Deutschen* (Ammon/Bickel/Lenz (Hg.) 2016) formulierten Bickel/Hofer/Suter (2015, S. 544) mehrere Desiderata, die ein Korpus als Datengrundlage für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen idealerweise erfüllen sollte:

1. „[...] die Textbasis [muss] gezielt nach national und regional zuordnenbaren Texten abgesucht [...] werden können."
2. „Das Korpus sollte [...] möglichst nur neuere und neuste standardsprachliche Texte enthalten."
3. „Das Korpus sollte groß genug sein, um auch bei selteneren Mehrwortverbindungen oder kleinräumigen Varianten aussagekräftige Treffermengen zu liefern."
4. „Hilfreich wäre zudem ein zuverlässiges Wortartentagging [...]."
5. „Schließlich wäre es wünschenswert, [...] dass mindestens absolute und relative Frequenzen einer Variante bzw. ihrer Formen in den Vollzentren des Deutschen automatisiert erhoben werden können."

„Ein linguistisches Korpus, das alle diese Wünsche erfüllt," stellten Bickel/Hofer/Suter (2015, S. 544) fest, „gibt es zur Zeit noch nicht." In Ermangelung eines solchen Korpus hat man bei der Erstellung der zweiten Auflage des Variantenwörterbuchs auf die wiso-Volltextdatenbank von GBI-Genios Deutsche Wirtschaftsdatenbank GmbH zurückgegriffen und für die deutschen Areale `D-nordwest`, `D-nordost`, `D-mittelwest`, `D-mittelost`, `D-südwest` und `D-südost` sowie die schweizerdeutschen und österreichischen Areale Teilkorpora erstellt, die jeweils mehrere Zeitungen umfassten.[1] Trotz der Aktualität, des Umfangs und der vielfältigen

**1** Für die schweizerdeutschen, österreichischen, liechtensteiner und rumänischen Areale wurden weitere Korpora herangezogen, die bei Ammon/Bickel/Lenz (Hg.) (2016, S. XV) aufgeführt sind.

Suchoperatoren der wiso-Datenbank vermissten Bickel/Hofer/Suter (2015, S. 546) insbesondere eine linguistische Annotation mit Wortarten-Tagging, Lemmatisierung und Eigennamenerkennung.

Neben den fünf Desiderata von Bickel/Hofer/Suter (2015) wäre ein weiteres, sechstes Desideratum anzuführen: die Beschränkung der Textauswahl auf Artikel aus Lokal- und Regionalteilen (bzw. Lokal- und Regionalressorts). Bei solchen Artikeln ist die Wahrscheinlichkeit, dass sie tatsächlich vor Ort entstanden sind, größer als bei Artikeln aus dem sogenannten Mantelteil, die häufig von überregionalen Zentralredaktionen oder Presseagenturen stammen. Dieser Ansatz wurde bei der Erstellung des Projektkorpus der *Variantengrammatik* verfolgt (Variantengrammatik des Standarddeutschen 2018), für das von Dezember 2011 bis Mai 2013 Artikel aus den Lokalteilen der Online-Ausgaben von 68 deutschsprachigen Tageszeitungen gecrawlt und linguistisch aufbereitet wurden. Insgesamt ergaben sich daraus knapp 600 Millionen laufende Wortformen. Für die areale Zuordnung wurde eine ähnliche Arealgliederung wie beim *Variantenwörterbuch* verwendet. (Zu den Einzelheiten vgl. Datenerhebung 2018.)

Ein siebtes Desideratum wäre schließlich, dass ein solches Korpus jedem an der Untersuchung diatopischer Variation im Deutschen Interessierten für eigene Recherchen zur Verfügung steht. Dies ist etwa beim Deutschen Referenzkorpus (DeReKo, Kupietz et al. 2018) des IDS der Fall, das über COSMAS II und KorAP nach einer Registrierung für wissenschaftliche und nicht-kommerzielle Zwecke allgemein nutzbar ist. Die umfangreichen Zeitungsquellen im DeReKo decken einen großen Teil des deutschsprachigen Raums ab. Allerdings sehen deren Metadaten weder eine areale Zuordnung vor, noch erlauben sie eine systematische Beschränkung auf Lokal- und Regionalteile.

# 3 Design und Arealkonzept des ZDL-Regionalkorpus

Das ZDL-Regionalkorpus entspricht den in Abschnitt 2 als Desiderata formulierten Kriterien:

1. Das ZDL-Regionalkorpus enthält Artikel aus deutschsprachigen Tageszeitungen, die mit Metadaten zu Land, Areal und Subareal versehen sind.
2. Das Korpus deckt den Zeitraum ab 1993 ab und wird monatlich aktualisiert.
3. Es umfasst aktuell 31,5 Mio. Artikel mit insgesamt 9,1 Mrd. Tokens (Stand: Mai 2022) und erlaubt somit auch Recherchen zu weniger frequenten Phänomenen.

4. Wie alle Korpora auf der DWDS-Plattform ist das ZDL-Regionalkorpus lemmatisiert und mit dem STTS-Tagset getaggt. Unter Bezug darauf lassen sich u. a. Abfragen formulieren, die Eigennamen aus der Suche ausschließen.
5. Die Benutzerschnittstelle des ZDL-Regionalkorpus bietet Recherchewerkzeuge zur Abfrage und kartographischen Visualisierung der Verteilung über Areale und Zeitungen an.
6. Durch die Beschränkung auf Lokal- und Regionalressorts werden Artikel von überregionalen Zentralredaktionen und Presseagenturen effektiv ausgeschlossen.
7. Das Korpus steht nicht nur den DWDS-Lexikographen, sondern allen registrierten Nutzern der DWDS-Plattform für eigene Recherchen zur Verfügung.

Für die Regionalangaben im DWDS und die arealen Metadaten im ZDL-Regionalkorpus wurde ein Arealkonzept erstellt, das die Arealgliederungen von *Variantenwörterbuch* und *Variantengrammatik* aufgreift. Dies betrifft insbesondere die Einteilung von Deutschland in sechs Areale mit den Bezeichnungen `D-Nordwest`, `D-Nordost`, `D-Mittelwest`, `D-Mittelost`, `D-Südwest` und `D-Südost`. Jedes Areal wurde in einem weiteren Schritt in Subareale wie `D-Südost (Franken)` oder `D-Südost (Altbayern)` unterteilt. Die Grenzziehung der Areale in kartographischen Darstellungen auf der DWDS-Plattform orientiert sich an der korpusbasierten Dialektgliederung bei Lameli (2013, S. 194), woraus sich in bestimmten Bereichen abweichende Arealgrenzen gegenüber der Arealkarte der *Variantengrammatik* (Datenerhebung 2018) ergeben.[2] Ein Überblick über die Arealgliederung in DWDS und ZDL-Regionalkorpus ist auf www.dwds.de/d/regionalangaben (Stand: 10.5.2022) verfügbar.

Aus jedem der sechs Areale in Deutschland wurden drei bis fünf Zeitungsquellen für das ZDL-Regionalkorpus ausgewählt. Tabelle 1 listet Areale und Zeitungen sowie die im ZDL-Regionalkorpus verfügbaren Zeiträume des Archivbestands auf. Die Arealkarte in Abbildung 1 lokalisiert die Zeitungsquellen an deren Haupterscheinungsort.

Die Rohdaten der Zeitungen werden mit Ausnahme der *Süddeutschen Zeitung* über GBI-Genios Deutsche Wirtschaftsdatenbank GmbH bezogen und an der Berliner Arbeitsstelle des ZDL als linguistisch annotiertes Korpus aufbereitet und monatlich aktualisiert. Wie erwähnt, gehen dabei nur Artikel aus Lokal- und Regionalressorts ins ZDL-Regionalkorpus ein. Aufgrund des unterschiedlichen Archiv-

**2** Diese Abweichungen betreffen u. a. die westfälischen Subareale, die im Arealkonzept von DWDS und ZDL-Regionalkorpus vollständig dem Areal `D-Nordwest` zugeordnet sind, sowie das saarländische Subareal, das hier Teil des Areals `D-Mittelwest` ist.

bestands bei Genios ergibt sich eine deutliche diachrone und areale Unausgewogenheit, wenn man den gesamten Abdeckungszeitraum ab 1993 betrachtet (Tab. 2); die areale Unausgewogenheit ist weniger ausgeprägt, wenn man sich auf den Zeitraum ab 2017 beschränkt, in dem aus allen Zeitungsquellen Daten im ZDL-Regionalkorpus vorhanden sind (Tab. 3). Dennoch ist es bei Arealvergleichen angebracht, sich auf relative Frequenzen statt nur auf absolute Trefferzahlen zu beziehen (vgl. Abschn. 4).

**Tab. 1:** Areale und Zeitungen im ZDL-Regionalkorpus

| Areal | Zeitung | Zeitraum |
|---|---|---|
| D-Nordwest | Hamburger Abendblatt | ab 1999 |
| | Kieler Nachrichten | ab 2017 |
| | Neue Osnabrücker Zeitung | ab 2012 |
| | Neue Westfälische | ab 2003 |
| D-Nordost | Berliner Morgenpost | ab 1999 |
| | Norddeutsche Neueste Nachrichten | ab 2012 |
| | Der Prignitzer | ab 2012 |
| | Schweriner Volkszeitung | ab 2004 |
| | Der Tagesspiegel | ab 2005 |
| D-Mittelwest | Aachener Zeitung | ab 2003 |
| | Allgemeine Zeitung (Mainz) | ab 2002 |
| | Frankfurter Rundschau | ab 1995 |
| | Rhein-Zeitung | ab 1997 |
| | Saarbrücker Zeitung | ab 1993 |
| D-Mittelost | Döbelner Allgemeine Zeitung | ab 2011 |
| | Dresdner Neueste Nachrichten | ab 2011 |
| | Leipziger Volkszeitung | ab 1997 |
| | Thüringer Allgemeine | ab 2000 |
| D-Südwest | Badische Zeitung | ab 2003 |
| | Reutlinger General-Anzeiger | ab 2007 |
| | Südkurier | ab 1999 |
| D-Südost | Fränkischer Tag | ab 2005 |
| | Landshuter Zeitung | ab 2014 |
| | Mittelbayerische | ab 2014 |
| | Münchner Merkur | ab 2016 |
| | Süddeutsche Zeitung | ab 2005 |


**Abb. 1:** Areale und Zeitungen im ZDL-Regionalkorpus

**Tab. 2:** Umfang des ZDL-Regionalkorpus im gesamten Abdeckungszeitraum (Stand: Mai 2022)

| Areal | Artikel | Tokens |
|---|---|---|
| D-Nordwest | 5,3 Mio. | 1,3 Mrd. |
| D-Nordost | 2,1 Mio. | 0,6 Mrd. |
| D-Mittelwest | 11,0 Mio. | 3,3 Mrd. |
| D-Mittelost | 5,6 Mio. | 1,6 Mrd. |
| D-Südwest | 3,4 Mio. | 1,0 Mrd. |
| D-Südost | 4,2 Mio. | 1,4 Mrd. |
| gesamt | 31,5 Mio. | 9,1 Mrd. |

**Tab. 3:** Umfang des ZDL-Regionalkorpus im Zeitraum ab 2017 (Stand: Mai 2022)

| Areal | Treffer | PPM |
|---|---|---|
| D-Nordwest | 1,5 Mio. | 0,4 Mrd. |
| D-Nordost | 0,5 Mio. | 0,1 Mrd. |
| D-Mittelwest | 1,4 Mio. | 0,5 Mrd. |
| D-Mittelost | 0,8 Mio. | 0,2 Mrd. |
| D-Südwest | 0,7 Mio. | 0,3 Mrd. |
| D-Südost | 2,5 Mio. | 0,8 Mrd. |
| gesamt | 7,5 Mio. | 2,3 Mio. |

# 4 Benutzerschnittstelle und Recherchewerkzeuge des ZDL-Regionalkorpus

Das ZDL-Regionalkorpus ist in die DWDS-Plattform eingebunden. Ein direkter Einstieg ist über die Korpus-Dokumentation auf www.dwds.de/d/korpora/regional (Stand: 10.5.2022) möglich. Gibt ein Nutzer auf dieser Seite eine Anfrage im Suchfeld ein, wird er auf die Anmeldeseite umgeleitet und bekommt nach erfolgreicher Anmeldung die Treffer im Suchinterface des ZDL-Regionalkorpus angezeigt. Dort kann die Anfrage nach Zeitraum und Arealen gefiltert sowie u. a. die Trefferssortierung eingestellt werden. Bei einer einfachen Abfrage wie `Fasching` werden per Default alle Treffer mit Formen des Lemmas „Fasching“ ausgegeben; dies ist die lexikographische Standardanwendung. Die DWDS-Abfragesprache erlaubt darüber hinaus komplexe Abfragen zu Wortformen, Phrasen, regulären Ausdrücken, Wortarten usw. (Näheres vgl. www.dwds.de/d/korpussuche, Stand: 10.5.2022).

Klickt man auf die Schaltfläche „Verteilung über Areale“, so erhält man eine tabellarische Frequenz-Übersicht, klassiert nach Arealen (Abb. 2). Neben absoluten Frequenzen werden relative Frequenzen als PPM-Werte (*parts per million*) ausgegeben: Treffer-Tokens pro Million Tokens im jeweiligen Areal im Abfrage-Zeitraum (im vorliegenden Beispiel: der Zeitraum 2017–2022, in dem aus allen Zeitungsquellen Daten vorhanden sind; vgl. Abschn. 3). Analoges gilt für die Schaltfläche „Verteilung über Zeitungen“, die eine nach Zeitungen klassierte Tabelle ausgibt (Abb. 3). Hier stehen die PPM-Werte für Treffer-Tokens pro Million Tokens in der jeweiligen Zeitung im Abfrage-Zeitraum. Die Schaltfläche „Karte anzeigen“ öffnet eine kartographische Visualisierung der PPM-Werte (vgl. Abb. 4).

Verteilung über Areale | Karte anzeigen | Tabelle als CSV

| Areal | Treffer | PPM | Anteil PPM |
|---|---|---|---|
| D-Südost | 21504 | 25,79 | 60,32 % |
| D-Mittelost | 2722 | 12,17 | 28,45 % |
| D-Nordost | 365 | 2,73 | 6,39 % |
| D-Mittelwest | 333 | 0,73 | 1,71 % |
| D-Südwest | 227 | 0,91 | 2,12 % |
| D-Nordwest | 169 | 0,43 | 1,00 % |

**Abb. 2:** Treffer und PPM-Werte für „Fasching“ pro Areal im Zeitraum ab 2017 (Stand: Mai 2022)

Verteilung über Zeitungen | Karte anzeigen | Tabelle als CSV

| Zeitung | Areal | Treffer | PPM |
|---|---|---|---|
| Mittelbayerische | D-Südost | 7832 | 33,11 |
| Münchner Merkur | D-Südost | 7175 | 19,71 |
| Landshuter Zeitung | D-Südost | 3614 | 37,27 |
| Fränkischer Tag | D-Südost | 2794 | 21,37 |
| Thüringer Allgemeine | D-Mittelost | 1779 | 14,72 |
| Leipziger Volkszeitung | D-Mittelost | 539 | 8,34 |
| Döbelner Allgemeine Zeitung | D-Mittelost | 293 | 19,51 |
| Schweriner Volkszeitung | D-Nordost | 243 | 4,01 |
| Saarbrücker Zeitung | D-Mittelwest | 174 | 2,11 |
| Reutlinger General-Anzeiger | D-Südwest | 124 | 3,09 |
| Dresdner Neueste Nachrichten | D-Mittelost | 111 | 4,78 |
| Südkurier | D-Südwest | 98 | 0,52 |
| Süddeutsche Zeitung | D-Südost | 89 | 16,40 |
| Kieler Nachrichten | D-Nordwest | 84 | 1,93 |
| Frankfurter Rundschau | D-Mittelwest | 65 | 1,41 |

**Abb. 3:** Treffer und PPM-Werte für „Fasching“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)


**Abb. 4:** PPM-Werte für „Fasching“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

# 5 Daten zur regionalen Variation aus dem ZDL-Regionalkorpus

In einer kleinen Fallstudie soll nun überprüft werden, inwieweit sich im ZDL-Regionalkorpus areale Verteilungen nachweisen lassen, die unabhängig für den *Atlas zur deutschen Alltagssprache* (AdA, Elspaß/Möller 2003–) erhoben wurden.

Als Fallbeispiel soll hier die AdA-Karte „Zeit vor dem Aschermittwoch“ mit den (Quasi-)Synonymen „Fasching“, „Karneval“ und „Fastnacht“ sowie dessen Varianten „Fasnacht“, „Fasenacht“ und „Fasnet“ dienen (www.atlas-alltagssprache.de/runde-2/f03/ (Stand: 1.3.2022), hier wiedergegeben als Abb. 5). Bei unseren Recherchen im ZDL-Regionalkorpus werden wir unter den Nicht-Standard-Varianten zusätzlich die beiden relativ frequenten Varianten „Carneval“ und „Fassenacht“ berücksichtigen.[3] Außerdem beschränken wir uns wieder auf den Zeitraum 2017–2022, in dem aus allen Zeitungsquellen Daten vorhanden sind (vgl. Abschn. 3). Aufgrund der unterschiedlichen Datentypen (Zeitungstexte im ZDL-Regionalkorpus vs. Sprechereinstellungen im AdA) wird keine vollständige Übereinstimmung zwischen den Daten im ZDL-Regionalkorpus und im AdA erwartet, wohl aber eine ähnliche Tendenz der arealen Verteilung.

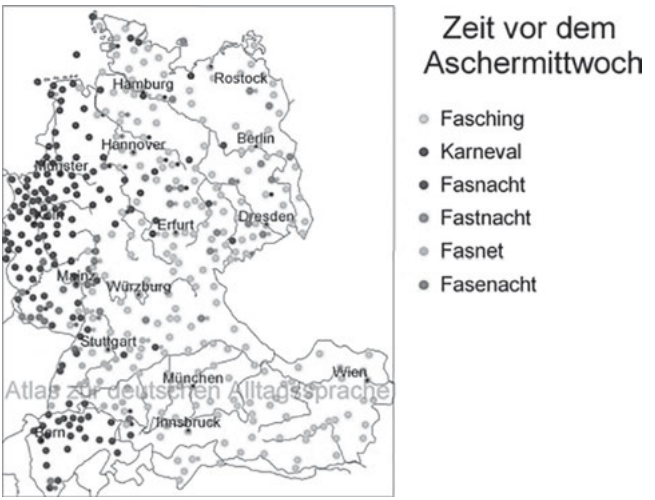


**Abb. 5:** Verteilung von „Fasching“, „Karneval“, „Fastnacht“, „Fasenacht“, „Fasnacht“ und „Fasnet“ im *Atlas zur deutschen Alltagssprache* (www.atlas-alltagssprache.de/runde-2/f03/, Stand: 1.3.2022)

**3** Die Varianten und ihre Frequenzverteilung wurden zunächst mit Hilfe der Abfragen `count(/^[CK]arn[aei]val$/) #by[$l]` und `count(/^F[ao]+[sß]+[te]?(n[ao]+cht|net)$/) #by[$l]` ermittelt. Im Sinne der Vergleichbarkeit mit der Lemma-Abfrage bei „Fasching“ (siehe oben Abschn. 4) wurde dann für die „Karneval“-Varianten und die „Fastnacht“-Varianten je eine disjunktive Lemma-Abfrage gestellt statt eine Abfrage mit regulären Ausdrücken, da sich letztere direkt auf die Token-Ebene statt auf die Lemmatisierung bezieht. Niedrigfrequente Varianten blieben bei den disjunktiven Abfragen unberücksichtigt.


**Abb. 6:** PPM-Werte für „Karneval“ und „Carneval“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)


**Abb. 7:** PPM-Werte für „Fastnacht“, „Fasenacht“, „Fassenacht“, „Fasnacht“ und „Fasnet“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

Aus der AdA-Karte ergibt sich die folgende großräumliche Verteilung. „Karneval“ ist die ortsübliche Variante im Nordwesten Deutschlands. Die „Fastnacht“-Varianten sind vor allem südlich der Mosel im Südwesten Deutschlands und in der deutschsprachigen Schweiz gebräuchlich. In anderen Regionen ist „Fasching“ die vorherrschende Variante; besonders eindeutig ist dies im Südosten Deutschlands sowie in Österreich.

Im ZDL-Regionalkorpus liegen die Maxima der PPM-Werte für die Varianten „Karneval“ und „Carneval“ im nordwestlichen Teil des Areals `D-Mittelwest`, und zwar bei der Aachener Zeitung und bei der Rhein-Zeitung, die in Koblenz und Umgebung erscheint (Abb. 6). Mittlere PPM-Werte für diese Varianten treten bei weiteren Zeitungen aus den Arealen `D-Mittelwest`, `D-Nordwest`, `D-Nordost` und `D-Mittelost` auf. Die Varianten „Fastnacht“, „Fasenacht“, „Fassenacht“, „Fasnacht“ und „Fasnet“ konzentrieren sich auf den südöstlichen Teil des Areals `D-Mittelwest` und das Areal `D-Südwest` (Abb. 7). Größere PPM-Werte für „Fasching“ treten vor allem in der östlichen Hälfte Deutschlands auf, mit einem eindeutigen Schwerpunkt auf den Arealen `D-Südost` und `D-Mittelost` (vgl. oben Abb. 4).

Alles in allem lässt sich festhalten, dass die generelle areale Verteilung der untersuchten Synonyme im AdA und im ZDL-Regionalkorpus durchaus vergleichbar ist. Dies betrifft insbesondere die Maxima der PPM-Werte im ZDL-Regionalkorpus, die durchweg in Regionen fallen, in denen im AdA dieselben Synonyme als ortsübliche Varianten erhoben wurden. Unterschiede zu den AdA-Daten treten im ZDL-Regionalkorpus vor allem bei niedrigeren relativen Frequenzen auf. Solche Divergenzen sind durchaus zu erwarten – zum einen, weil in den Artikeln ein und derselben Zeitungsquelle im Allgemeinen verschiedene Synonyme vorkommen, und zum anderen aufgrund von Marketing-Schöpfungen wie dem Berliner ‚Karneval der Kulturen‘, die bewusst ortsuntypische Verwendungen assoziieren.

## 6 Lexikographische Praxis

In diesem Abschnitt soll auf die Nutzung des ZDL-Regionalkorpus für die Überarbeitung der Regionalangaben in den lexikographischen Substanzen des DWDS eingegangen werden. Diese basieren zu großen Teilen auf zwei Wörterbüchern: dem *Wörterbuch der deutschen Gegenwartssprache* (WDG, Klappenbach/Steinitz 1961–1977) und dem *Großen Wörterbuch der deutschen Sprache* (GWDS 1999).

Die regionalsprachlichen Markierungen in beiden Wörterbüchern soll im Rahmen des ZDL anhand der Daten des ZDL-Regionalkorpus geprüft und gemäß des oben beschriebenen Arealkonzepts überarbeitet werden. Dies betrifft etwa 6500 Wörterbuchartikel in den beiden Substanzen.

Aus lexikographischer Sicht unterscheiden wir Sachspezifika und Regionalismen. Unter Sachspezifika fassen wir Wörter, die einer Region zuzuordnen sind, aber auch außerhalb dieser Region so benannt werden, z. B. „Rösti“ oder „Printe“. Als Regionalismen hingegen sehen wir regionalspezifisch Wörter an, für die es Synonyme in anderen Regionen gibt, z. B. „Karneval“, „Fasching“, „Fastnacht“ oder „Fasnacht“.

- Im Wörterbucheintrag werden Sachspezifika über Definitionszusätze markiert: z. B. „Rösti“ (‚aus der Schweizer Küche stammend‘) oder „Printe“ (‚aus dem Rheinland stammend‘, siehe Abb. 8).
- Regionalismen hingegen werden mit Areal- und ggf. Subarealangaben versehen, z. B. „Karneval“, „Fasching, „Fas(t)nacht“ (vgl. Abb. 9 bzw. die entsprechenden Wörterbuchartikel[4]). In den Wörterbuchartikeln werden diese Wörter auf der Basis der Karten den jeweiligen Arealen zugeordnet: Bei „Karneval“ ist das „besonders `D-Mittelwest`“, bei „Fasching“ „besonders `D-Südost`, oft `D-Mittelost`, `A`“, bei „Fastnacht“ „besonders `D-Mittelwest`, gelegentlich `D-Südwest`“ und bei „Fasnacht“ „`D-Südwest`, `CH`“. Die gegenüber „Fastnacht“ auffällig höheren Verwendungen der Schreibweise „Fasnacht“ in diesen beiden Arealen in verschiedenen Presseerzeugnissen in nicht-dialektalen textuellen Umgebungen spricht aus unserer Sicht dafür, diese Schreibweise auch in einem allgemeinsprachlichen Wörterbuch als Regionalismus aufzunehmen.


**Abb. 8:** DWDS-Artikel zum Lemma „Printe“

4 Die genannten DWDS-Artikel sind unter www.dwds.de/wb/Karneval, www.dwds.de/wb/Fasching, www.dwds.de/wb/Fastnacht und www.dwds.de/wb/Fasnacht zu finden (Stand: 10.5.2022).

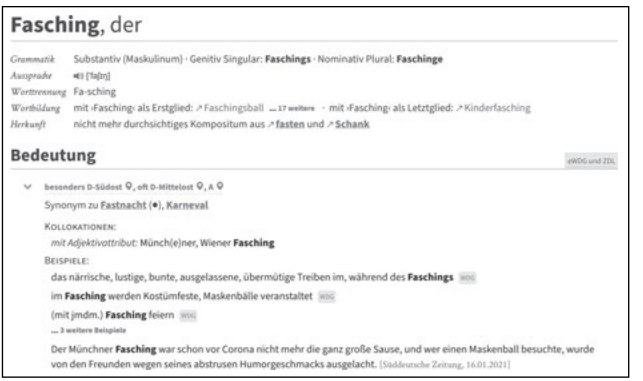


**Abb. 9:** DWDS-Artikel zum Lemma „Fasching“

In diesen Beispielen bietet das ZDL-Regionalkorpus ausreichende Belegzahlen und regionale Präferenzen an, um die Überarbeitung auf der Grundlage der Korpusfrequenzen vorzunehmen. Nicht immer ist dies jedoch tatsächlich der Fall. Wie bei der Korpusanalyse generell, so lauern auch im Falle der lexikographischen Bewertung der Daten des ZDL-Regionalkorpus einige Fallstricke. Neben den bekannten Fällen, wie die Homographie von Eigennamen und Appellativa, die zu unplausiblen Frequenzverteilungen der regionalen Verteilung führen können, ist auch die unzureichende Beleglage ein Problem. Dies ist auf den ersten Blick erstaunlich, da das ZDL-Regionalkorpus mit über 9 Mrd. Tokens hinreichend groß erscheint. Im Unterschied zu Wörtern des Standardwortschatzes muss bei Regionalismen nicht nur darauf geachtet werden, ob die Gesamtfrequenz hinreichend groß ist; auch die Verteilung über die Areale muss signifikante Unterschiede aufweisen. Dass es sich hier nicht nur um Randphänomene handelt, zeigen die folgenden drei Beispiele: das Wort „Gschaftlhuber“ (74 Treffer) lässt noch eine Aussage darüber zu, dass dieses Wort wohl dem Areal `D-Südost` zuzuordnen ist. Bei der nominalen Ableitung „Gschaftlhuberei“ (18 Treffer über alle Areale) ist aufgrund der Trefferanzahl eine fundierte Bewertung nur bedingt möglich.

Neben der zu geringen Frequenz gibt es auch gelegentlich das Phänomen der schwierigen Interpretierbarkeit von regionalen Frequenzunterschieden. Diese müssen nicht immer auf einen Regionalismus zurückzuführen sein. Ein Beispiel hierfür ist „Knecht Rup(p)recht“. Hier ist der Anteil in der `D-Südwest` signifikant höher als im Rest (42 PPM in `D-Südwest` gegenüber ca. 10 PPM in den anderen Arealen in `D`). Sieht man jedoch in den Einzequellen nach, so stellt man fest, dass sich der hohe Anteil in `D-Südwest` nur auf eine einzige Quelle zurückführen lässt: den „Südkurier“.

Als Zwischenfazit nach bislang etwa 2.000 für das DWDS überarbeiteten und publizierten Regionalismen und Sachspezifika (dies entspricht etwa einem Drittel der in WDG und GWDS diatopisch markierten Stichwörter) lässt sich festhalten, dass das ZDL-Regionalkorpus in den meisten Fällen eine ausreichende Grundlage für die lexikographische Entscheidung bereitstellt, zumindest dann, wenn die Korpusfrequenzen im Zuge der Arbeit eine lexikographische Interpretation erfahren. Die Grenzfälle bezüglich der Seltenheit verbleiben aber auch bei einem Korpus von 9 Mrd. Textwörtern.

Gegenwärtig besteht eine weitere Beschränkung des ZDL-Regionalkorpus darin, dass es nur Quellen aus Deutschland enthält. Wir haben aus diesem Grund ein zweites Korpus angelegt, welches das ZDL-Regionalkorpus vor allem in geographischer Hinsicht erweitert: das Webmonitor-Korpus, auf das wir im folgenden Abschnitt eingehen wollen.

# 7 Ergänzungen aus dem Web

## 7.1 Hintergrund

Der Bestand des ZDL-Regionalkorpus ist vertraglich abgesichert, deswegen wird seine Zusammensetzung auch von Fragen der Lizenzierung geprägt. Andere Länder oder bestimmte Subareale können nicht ohne Weiteres einbezogen werden, obwohl eine größere Quellenvielfalt sowohl im Hinblick auf die qualitative (u. a. lexikographische) Analyse als auch auf die quantitative Aussagekraft der Korpustreffer wünschenswert wäre.

Vor diesem Hintergrund erscheint eine Ergänzung um Internetquellen sinnvoll, deren wissenschaftliche Nutzung nach entsprechenden gesetzlichen Änderungen (in Deutschland: UrhWissG § 60) ins Blickfeld gerät. Auch wenn bereits auf der DWDS-Plattform existierende, breitgefächerte Blog- und Nachrichtenkorpora aus dem Web über die nötige Größe verfügen, um diverse Fragen zur Sprachnutzung zu beantworten, waren sie nicht unmittelbar mit den Zeitungsquellen des ZDL-Regionalkorpus vergleichbar. Insbesondere wiesen sie weder eine auf Zeitungsartikel fokussierte Textgrundlage noch adäquate Metadaten auf.

Zu diesem Zweck wurde ein Monitorkorpus aus Internetquellen adaptiert. Die Methodik, die der generischen Entdeckung und Erschließung von Online-Texten auf der DWDS-Plattform zugrunde liegt, ist reproduzierbar (Barbaresi 2021) und kann an verschiedene Anforderungen angepasst werden. Sie umfasst die Hauptphasen der Datenerhebung im Sinne der *Variantengrammatik*: Datenakquisition, Datenbereinigung und Dubletten-Erkennung (Datenerhebung 2018).

Im Folgenden wird zunächst das Webmonitor-Korpus näher beschrieben; danach werden einige Nutzungsszenarien für Analysen auf lexikalischer Ebene exemplarisch vorgeführt.

## 7.2 Das Webmonitor-Korpus

Das Webmonitor-Korpus wurde Anfang 2021 angelegt, um ein Webkorpus auf der DWDS-Plattform aktuell zu halten und gleichzeitig besonders wertvolle Quellen zu kuratieren und zur Verfügung zu stellen.

Es besteht aus einem allgemeinen Korpus aus prominenten Quellen, das täglich aktualisiert wird, indem Web-Feeds gesammelt werden und entsprechende Seiten heruntergeladen, verarbeitet und indiziert werden. Bemerkenswerte Unterschiede zu vergleichbaren Unternehmungen von Biemann et al. (2007) oder Minocha/Reddy/Kilgarriff (2014) betreffen den Auswahlprozess: Erstens zählt die in Webseitenbesuchen geschätzte Größe der Leserschaft als Aufnahmekriterium, zweitens wird eine gewisse thematische Balance zwischen den Quellen in Betracht gezogen, und drittens sind die Quellen nicht nur journalistischer Natur.

Derzeit umfasst das Webmonitor-Korpus 1,7 Mrd. Tokens aus über 500 Quellen, 3 bis 4 Mio. Tokens kommen täglich neu hinzu (Stand: Mai 2022). Dabei stehen größere Nachrichtenportale sowie die überregionale und regionale Presse im Fokus. Viele der meistgelesenen Internetseiten im deutschen Sprachraum zählen dazu, auch die Regenbogenpresse sowie Gratis- und Boulevardzeitungen sind im Korpus recherchierbar. Weitere spezialisierte Webseiten mit einer breiten Leserschaft ergänzen die Sammlung, darunter Nachrichtenseiten zu diversen Berufsgruppen und Hobbys. Ein weiterer Schwerpunkt sind offizielle Webpräsenzen von Behörden (unter anderem Seiten von Ministerien, Bundesländern, Großstädten und Kantonen) und prominenten Nichtregierungsorganisationen (NGOs).

Diese Daten sind insbesondere für jüngste Entwicklungen relevant; damit können die DWDS-Lexikographen neueste Trends bei der Formulierung von Definitionen und der Auswahl von Beispielsätzen berücksichtigen. Außerdem lassen sich daraus Trendwörter auf der Basis von Frequenzinformationen ermitteln, mit deren Hilfe Kandidaten für die Aufnahme ins Wörterbuch oder die Bearbeitung von existierenden Artikeln bestimmt werden können.

## 7.3 Regionalteile

Ein durch areale Metadaten ausgezeichnetes Subkorpus mit regionalen Quellen wurde Anfang 2022 zum Webmonitor-Korpus hinzugefügt, und zwar einerseits

durch die nähere Spezifizierung bereits erfasster Zeitungen und andererseits durch zusätzlich aufgenommene regionale Quellen.

Ungefähr 90 Quellen, die explizite Regionalteile aufweisen, werden nach dem ZDL-Arealkonzept (vgl. Abschn. 3) in linguistisch relevante Areale unterteilt, die je nach Bedarf und Verfügbarkeit gefüllt werden. Damit können Regionalismen im Sinne der Regionalangaben im DWDS in Deutschland, Österreich, der Schweiz, Italien (Südtirol), Belgien (Ostbelgien), Luxemburg, und Liechtenstein gezielt abgefragt werden. Diese Ergänzung liefert auch zusätzliche Ortspunkte bei bereits existierenden Arealen und Subarealen (z. B. bei `D-Südwest`).

Dokumente ohne areale Metadaten können anhand von anderen Metadaten gezielt abgefragt werden. Erstens werden lokale Online-Zeitungen, die keinem Areal zugeordnet werden können (z. B. in Mallorca oder Thailand), nicht berücksichtigt. Zweitens weichen bestimmte Seiten im Hinblick auf ihre Textgenres von dem Rest ab. So ist eine Kantonsseite örtlich relevant und kann einzeln oder zusammen mit anderen Quellen aus der Schweiz anhand von URL-Merkmalen (hier: `.ch`) gezielt abfragt werden.

Im Vergleich mit dem ZDL-Regionalkorpus ist die Quellenlage dynamisch. Bei Bedarf werden die Quellen im Webmonitor-Korpus angepasst und ergänzt, beispielsweise bei einem Ausfall oder falls weitere relevante Seiten gefunden werden. Dank der beschriebenen Kriterien und der Möglichkeit, relative Frequenzen in Form von PPM-Werten zu erheben bleiben regional relevante Ergebnisse vergleichbar, während eine mögliche Erweiterung zusätzliche interessante Belege für die qualitative Arbeit liefert. Insgesamt sollten das Korpus und der regionale Teil durch organisches Wachstum und Erweiterungen allmählich an chronologischer Tiefe gewinnen.

#### Beispiel 1: Aggregierte Informationen, Komposita auf „Sackerl"-Basis

Die Korpora zusammen ermöglichen gruppierte statistische Untersuchungen auf der Basis von Suchergebnissen, hier einem besonders in Österreich attestierten *-erl*-Diminutiv (Schwaiger et al. 2019). Alle auf *-sackerl* endenden Formen können aggregiert werden, die häufigsten sind (Stand: Mai 2022): *Plastiksackerl* (93 Vorkommen), *Papiersackerl* (20), *Überraschungssackerl* (18), *Einkaufssackerl* (13), *Nikolaussackerl* (15), *Stoffsackerl* (11), *Biosackerl* und *Nikolosackerl* (10), *Blühwiesen-Samensackerl*, *Frühstückssackerl* und *Geschenksackerl* (je 7), *Startersackerl* (6), *Mistsackerl* und *Teesackerl* (je 5), *Jausensackerl* (4).

Die PPM-Werte zeigen, dass diese Begriffe in den Daten trotz der unterschiedlichen Korpusgrößen anders akzentuiert werden: 308 Treffer insgesamt im Webmonitor-Korpus (184 PPM) und 158 Treffer insgesamt im ZDL-Regionalkorpus (17 PPM). Dieser Unterschied ist auf die fehlenden österreichischen Daten im ZDL-Regionalkorpus zurückzuführen.

**Beispiel 2: Tabellarische Informationen: das Wort „Lokal“**

Die geographische Verteilung von Treffern kann in Form einer Tabelle ausgegeben werden. Mit dieser Übersicht können auch lokale Unterschiede in Arealen festgestellt werden, die nicht im Regionalkorpus enthalten sind.

Abbildung 10 zeigt aggregierte Ergebnisse für das Lemma „Lokal“ (als Nomen und mit seinen möglichen Formen wie z. B. *Lokals*) in allen im Webmonitor-Korpus verzeichneten Arealen und anhand der PPM-Werte sortiert. So erscheint an erster Stelle das Areal, wo das Wort am häufigsten vorkommt, hier `CH`. Insgesamt fallen die relativen Frequenzen (PPM-Werte) in den Arealen `CH`, `A`, `D-Südost`, `STIR` und `LUX` auf, während `D-Mittelost` verhältnismäßig die niedrigste Häufigkeit aufweist. Diese Diskrepanz kann durch die Verwendung des Wortes/Lemmas für ‚Bar‘ oder ‚Gaststätte‘ im Süden und Südwesten erklärt werden, wohingegen er in den anderen Arealen von anderen Wörtern verdrängt wird. Die PPM-Werte liefern auch graduelle Informationen zu diatopisch erfassbaren Phänomenen, die wie eben beschrieben als relative Frequenzen aufgelistet werden oder prozentual hinsichtlich der ganzen Treffermenge (4. Spalte).

| Areal | Treffer | PPM | Anteil PPM |
|---|---|---|---|
| CH | 156 | 69,50 | 19,04 % |
| A | 1640 | 52,63 | 14,42 % |
| D-Südost | 439 | 42,03 | 11,52 % |
| STIR | 108 | 38,35 | 10,51 % |
| LUX | 38 | 37,69 | 10,33 % |
| LIE | 4 | 24,63 | 6,75 % |
| D-Mittelwest | 372 | 23,77 | 6,51 % |
| D-Südwest | 247 | 18,85 | 5,16 % |
| D-Nordost | 42 | 17,25 | 4,73 % |
| D-Nordwest | 208 | 16,35 | 4,48 % |
| BELG | 4 | 15,46 | 4,24 % |
| D-Mittelost | 47 | 8,48 | 2,32 % |

**Abb. 10:** Verteilung über Areale für das Lemma „Lokal“ (Stand: Mai 2022)

### 7.4 Beispiele aus Abschnitt 5

Bei dem jetzigen Stand liefert das Webmonitor-Korpus weitere Informationen zu den in Abschnitt 5 besprochenen Beispielen: „Fasching“ ist in den Arealen `D-Südost`, `D-Südwest` und `A` prominent; „Karneval“ ist in `BELG` besonders prominent und in `D-Mittelwest`, `D-Nordwest` und `LUX` gut vertreten. Die verschiedenen orthographischen Varianten für „Fastnacht“ sind besonders in `CH` zu finden und ansonsten auch in `LIE` und `D-Südwest` gut vertreten. Auch wenn mit etwas mehr Datenrücklauf noch feinere arealtypische Gewichtungen möglich sein werden, zeigt sich bereits jetzt, dass die zusätzlichen Areale und Länder nützlich für die Analyse sind.

## 8 Ausblick: Grenzen der beschriebenen Korpora

Im ZDL-Kontext haben sich das ZDL-Regionalkorpus und das Webmonitor-Korpus grundsätzlich als Datengrundlage für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache bewährt. Dennoch wäre es unangebracht, diesen Beitrag zu beenden, ohne auf die Grenzen ihrer Leistungsfähigkeit für variationslinguistische Untersuchungen zum Deutschen zu verweisen.

Beide Korpora enthalten Pressetexte aus Print- oder Onlinemedien. Damit fehlen wichtige Textsorten für die Untersuchung diatopischer Variation. Eine interessante Ergänzung wären hier Sammlungen regionaler Belletristik von der Mundartdichtung bis zu Regionalkrimis.[5]

Mit dem Medium der Quellen geht einher, dass gesprochene Sprache in den Korpora kaum repräsentiert ist – von Interviews und Zitaten mündlicher Rede einmal abgesehen. Insbesondere fehlen weitgehend Äußerungen (basis-)dialektaler Art, die allenfalls als Kuriosum in Ressorts wie „Uff Hunsrigga Platt“ (*Rhein-Zeitung*) erscheinen.[6]

In Ermangelung von Standards geschieht der Prozess der Verschriftlichung bei dialektal verwendeten Ausdrücken und Wörtern oft unterschiedlich, so dass in den Textkorpora erheblich voneinander abweichende orthographische Formen zu finden sind, wie oben anhand des Beispiels der „Fastnacht“-Varianten gezeigt. Falls solche Varianten zahlreich oder schwierig zu ermitteln sind, ist ein regulä-

**5** Regionalkrimis sind Teil des DeReKo-Korpus `lit-pub` (Belletristik/Trivialliteratur).

**6** Für genuin dialektologische Daten sei unter anderem auf die Ressourcen des Projekts Regionalsprache.de (REDE) verwiesen.

rer Ausdruck das Mittel der Wahl (vgl. den regulären Ausdruck in Fußnote 3, mit dem im gleichen Abfrage-Zeitraum ca. 5% mehr Treffer als mit der disjunktiven Lemma-Abfrage der „Fastnacht“-Varianten gefunden werden, Stand: Mai 2022).

Ein anderes Beispiel bezieht sich auf regionale Formen für das Wort „Kartoffel“, die sich mit dem folgenden regulären Ausdrück suchen lassen: `/[gk]r[ou]m+b[ei]+r[ae]?(\W|$)/i`. Unter den 198 Treffern im Webmonitor-Korpus (Abb. 11, Stand: Mai 2022) können die folgenden Formen attestiert werden: *Grumbeere* (Pfalz), *Grombier*/*Grombira* (Stuttgart), *Krumbiere* (Schwarzwald), *Krumbeer* (Trier). Diese Schreibweisen reflektieren gewisse Merkmale unterschiedlicher Aussprache, die qualitativ eingeordnet werden könnten. Ob mit dem obigen regulären Ausdruck alle möglichen Formen zu finden sind, ist nicht sicher. Außerdem prägt die Öffentlichkeitsarbeit zur ‚Pfälzer Grummbeere‘ die Treffer entscheidend.


**Abb. 11:** zufällige Trefferansicht, regional verwendete Alternative für „Kartoffel“

In den Print-Quellen des ZDL-Regionalkorpus werden mit der größeren zeitlichen Tiefe verhältnismäßig mehr Treffer gefunden. Die tokenstarke *Rhein-Zeitung* liefert weitere Belege für das Areal `D-Mittelwest`. Auch wenn viele mögliche Formen der Verschriftlichung gefunden werden, können Regionalquellen (print und

online) stark von außerlinguistischen Faktoren wie Werbekampagnen oder politischen Ereignissen beeinflusst sein.

Als Korpora der Gegenwart beleuchten das ZDL-Regionalkorpus und das Webmonitor-Korpus den aktuellen Stand der deutschen Sprache und ihrer standardsprachlichen Varietäten. Trotz regelmäßiger Aktualisierung eignen sich die Korpora aufgrund der noch zu geringen diachronen Tiefe kaum für Untersuchungen zeitlicher Verläufe. Zudem sind historisch belegte Varianten nicht in den Korpora enthalten, wenn sie aktuell nicht attestierbar sind – insbesondere dann, wenn wie beispielsweise im Elsass Zeitungen und Online-Nachrichten mit einer vergleichbaren Leserschaft nicht mehr auf Deutsch erscheinen.

## Literatur

Ammon, Ulrich/Bickel, Hans/Lenz, Alexandra N. (Hg.) (2016): Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. 2., völlig neu bearb. u. erw. Aufl. Berlin/Boston: De Gruyter.

Barbaresi, Adrien (2021): Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In: Ji, Heng/Park, Jong C./ Xia, Rui (Hg.): Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Stroudsburg, S. 122–131.

Bickel, Hans/Hofer, Lorenz/Suter, Sandra (2015): Variantenwörterbuch des Deutschen (VWB) – NEU: Dynamik der deutschen Standardvariation aus lexikografischer Sicht. In: Kehrein, Roland/Lameli, Alfred/Rabanus, Stefan (Hg.): Regionale Variation des Deutschen: Projekte und Perspektiven. Berlin/Boston: De Gruyter, S. 541–562.

Biemann, Chris/Heyer, Gerhard/Quasthoff, Uwe/Richter, Matthias (2007): The Leipzig Corpora Collection – Monolingual corpora of standard size. In: Proceedings of Corpus Linguistics conference, University of Birmingham, 27–30 July 2007. Birmingham.

Clear, Jeremy (1987): Trawling the language: monitor corpora. In: Snell-Hornby, Mary (Hg.): Züri-LEX '86 Proceedings: Papers read at the Euralex International Congress, University of Zürich, 9–14 September 1986. Tübingen: Francke.

Datenerhebung (2018). In: Variantengrammatik des Standarddeutschen: Ein Online-Nachschlagewerk. Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler. http://mediawiki.ids-mannheim.de/VarGra/index.php/Datenerhebung (Stand: 10.5.2022).

Elspaß, Stephan/Möller, Robert (2003–): Atlas zur deutschen Alltagssprache (AdA). www.atlas-alltagssprache.de (Stand: 10.5.2022).

Geyken, Alexander/Barbaresi, Adrien/Didakowski, Jörg/Jurish, Bryan/Wiegand, Frank/Lemnitzer, Lothar (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache" (DWDS). In: Zeitschrift für germanistische Linguistik 45, 2, S. 327–344.

GWDS (1999) = Duden (1999): Das große Wörterbuch der deutschen Sprache. 10 Bde. 3., völlig neu bearb. und erw. Aufl. Mannheim u. a.: Dudenverlag.

Klappenbach, Ruth/Steinitz, Wolfgang (1961 –1977): Wörterbuch der deutschen Gegenwartssprache. 6 Bde. Berlin: Akademie-Verlag.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New developments – new opportunities. In: Calzolari, Nicoletta/Coukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7–12 May 2018, Miyazaki, Japan. Paris: European Language Resources Association (ELRA), S. 4353–4360.

Lameli, Alfred (2013): Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland. (= Linguistik – Impulse & Tendenzen 54). Berlin/Boston: De Gruyter.

Minocha, Akshay/Reddy, Siva/Kilgarriff, Adam (2014): Feed corpus: an ever growing up-to-date corpus. In: Evert, Stefan/Stemle, Egon/Rayson, Paul (Hg.): Proceedings of the 8th Web as Corpus Workshop, ACL SIGWAC. Lancaster: WAC-8 Organising Committee, S. 1–4.

Schwaiger, Sonia/Barbaresi, Adrien/Korecky-Kröll, Katharina/Ransmayr, Julia/Dressler, Wolfgang (2019): Diminutivvariation in österreichischen elektronischen Corpora. In: Bülow, Lars/Fischer, Ann Kathrin/Herbert, Kristina (Hg.): Dimensionen des sprachlichen Raums: Variation – Mehrsprachigkeit – Konzeptualisierung. (= Schriften zur deutschen Sprache in Österreich 45). Berlin u. a.: Lang, S. 147–162.

Sinclair, John (1982): Reflections on computer corpora in English language research. In: Johansson, Stig (Hg.): Computer corpora in English language research. Bergen: Norwegian Computing Centre for the Humanities, S. 1–6.

Variantengrammatik des Standarddeutschen: Ein Online-Nachschlagewerk (2018). Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler. http://mediawiki.ids-mannheim.de/VarGra/ (Stand: 10.5.2022).