Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS)

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand, Lothar Lemnitzer

FINAL DRAFT – publication available at https://doi.org/10.1515/zgl-2017-0017

Linguistik im Internet, Zeitschrift für Germanistische Linguistik 2017; 45(2)

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand und Lothar Lemnitzer

1 Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS)

1. Einleitung

Ziel des an der Berlin-Brandenburgischen Akademie der Wissenschaften beheimateten Vorhabens „Digitales Wörterbuch der deutschen Sprache“ (DWDS) ist die Schaffung eines Digitalen Lexikalischen Systems – eines umfassenden, jedem Benutzer über das Internet zugänglichen Wortinformationssystems, das Auskunft über den deutschen Wortschatz in Vergangenheit und Gegenwart gibt. Das Vorhaben ist auf achtzehn Jahre angelegt (2007–2024, vgl. Klein & Geyken 2010)[1].

Für die Erarbeitung dieses Wörterbuches stehen den Lexikographinnen und Lexikographen verschiedene Korpora des gegenwartssprachlichen Deutsch zur Verfügung. Diese wurden in den vergangenen 15 Jahren im Rahmen des Vorhabens aufgebaut bzw. von Verlagen eingeworben und decken den Gegenstand des Wörterbuchs, d. h. den Wortschatz des Deutschen im 20. und im frühen 21. Jahrhundert, ab. Damit schließen diese Korpora an das ebenfalls an der BBAW beheimatete Projekt „Deutsches Textarchiv“ (DTA) an, in dem ein Referenzkorpus und verschiedene Ergänzungskorpora des Deutschen für den Zeitraum von ca. 1600 bis 1900 digital erfasst, aufbereitet und für die Forschung zugänglich gemacht wurden[2].

Die Kernaufgabe der Projektgruppe des DWDS besteht darin, den in den Korpora enthaltenen Wortschatz lexikographisch und korpusbasiert zu beschreiben. Es erschien uns aber von Beginn des Projekts an sinnvoll, die Daten, die die Basis dieser lexikographischen Arbeiten bilden, auch allgemein der empirischen linguistischen Forschung zum Deutschen in Geschichte und Gegenwart zur Verfügung zu stellen. Durch den Zugriff auf die durch rechtliche Vereinbarungen geklärten Teile der Wörterbuchbasis ist es somit den Benutzerinnen und Benutzern unseres Wörterbuchs jederzeit möglich, die lexikographischen Angaben in unserem Wörterbuch zu prüfen, aber auch eigene korpusbasierte Recherchen auf den verschiedenen Korpora des DWDS durchzuführen. Wir halten es aber für wichtig zu betonen, dass sich die Akquisition und Bereitstellung von Texten primär an den lexikographischen Erfordernissen des Projekts ausrichtet.

---

1 s. https://www.dwds.de/.

2 http://www.deutschestextarchiv.de/. Eine ausführliche Beschreibung dieses Korpus ist für die übernächste Ausgabe der ZGL-Rubrik „Linguistik im Internet“ vorgesehen.

Wir werden im Folgenden die verschiedenen Korpora des DWDS darstellen (Abschnitt 2). In Abschnitt 3 gehen wir auf technische und rechtliche Aspekte des Korpusmanagements ein. In Abschnitt 4 stellen wir einige Analysewerkzeuge dar, die einerseits als Hilfsmittel für die tägliche lexikographische Arbeit dienen, andererseits auch für die Nutzerinnen und Nutzer der DWDS-Plattform zur Verfügung stehen.

2. Umfang und Struktur der DWDS-Korpora

Die Sammlung von Korpora im DWDS umfasst derzeit (Stand: Mai 2017) ca. 12,6 Milliarden Textwörter in etwa 750 Millionen Sätzen. Sie gliedert sich in drei Bereiche: a) Referenzkorpora, b) Zeitungskorpora und c) Spezialkorpora. Nicht alle Texte sind frei verfügbar bzw. öffentlich recherchierbar. Dies hängt mit den Nutzungs- und Wiederverwertungsrechten zusammen, die dem DWDS von den Inhabern der Urheber- und Verwertungsrechte an den Texten eingeräumt wurden. Wir werden in Abschnitt 3 näher darauf eingehen. In der folgenden Beschreibung wird der Umfang jedes Korpus in Tokens und Anzahl der Dokumente angegeben.[3]

2.1 Referenzkorpora

In die Rechercheplattform des DWDS sind drei Referenzkorpora eingebunden, die insgesamt einen Zeitraum von über 400 Jahren von ca. 1600 bis heute umspannen.

2.1.1 Korpora des Deutschen Textarchivs (DTA)

Ziel des von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft (DFG) geförderten DTA war die Erstellung eines disziplinen- und gattungsübergreifenden Grundbestands deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis etwa 1900. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern der BBAW kommentierten und ergänzten, umfangreichen Bibliographie. Aus dieser wurde ein nach Textsorten und Disziplinen ausgewogenes Textkorpus zusammengestellt, das als Grundlage für ein Referenzkorpus zur Entwicklung der neuhochdeutschen Sprache dient. Das nach diesen Kriterien zusammengestellte DTA-Kernkorpus wird kontinuierlich erweitert. Das zeitlich und nach Textsorten ausgewogene DTA-Kernkorpus umfasst 100 Millionen Textwörtern (120 Millionen Tokens) in 1 500 Werken. Zusammen mit dem DTA-Ergänzungskorpus umfasst das Korpus des deutschen Textarchivs ca. 200 Millionen Tokens. Das gesamte DTA-Korpus steht unter einer offenen Lizenz und damit nachnutzbar zum Download auf der DTA-Webseite zur Verfügung.

2.1.2 DWDS-Kernkorpus des 20. Jahrhunderts

Das *DWDS-Kernkorpus des 20. Jahrhunderts* ist ein nach Textsorten und zeitlich über das gesamte Jahrhundert ausgewogenes Korpus im Umfang von 100 Millionen Textwörtern in 80 000 Dokumenten. Vom Korpusumfang ist es damit vergleichbar mit dem DTA-Kernkorpus. Es besteht aus den vier Textsortenbereichen Belletristik, Zeitung,

3 Aktuell(er)e Zahlen können jederzeit unter https://www.dwds.de/r/stat abgefragt werden.

Wissenschaft, Gebrauchsliteratur. Das Verhältnis der Textsortbereiche in der auf der Webseite veröffentlichten Version des DWDS-Kernkorpus ist wie folgt: a) Belletristik: 28,42 %; b) Zeitung: 27,36 %; c) Wissenschaft: 23,15 %; d) Gebrauchsliteratur: 21,05 %. Das Verteilungsmuster wurde, so weit es möglich war, auch über die einzelnen Dekaden des Jahrhunderts angewendet. Zum Aufbau des Korpus vgl. Geyken (2007).

2.1.3 DWDS-Kernkorpus des 21. Jahrhunderts

Das *DWDS-Kernkorpus des 21. Jahrhunderts* ist ein zeitlich und nach den in 2.1.2 genannten Textsorten differenziertes, derzeit aber noch nicht ausgewogenes Korpus. Das Korpus wird fortlaufend um Texte aller Textsorten ergänzt, sobald ausreichende Rechte für die Anzeige dieser Texte vorliegen. Das Kernkorpus des 21. Jahrhundert umfasst derzeit 15 Millionen Token in 12 000 Dokumenten.

Alle drei Referenzkorpora sind über eine linguistische Suchmaschine recherchierbar und werden für einige DWDS-Anwendungen, z. B. die Wortverlaufskurve (s. Abschnitt 4), verwendet.

2.2 Zeitungskorpora

Etwa 20 % der Zeitungskorpora des DWDS sind frei, d. h. ohne Anmeldung zugänglich, weitere 15 % können erst nach (kostenfreier) Anmeldung angezeigt werden. Für den verbleibenden Teil gilt, dass die Nutzung der Texte auf die lexikographischen Arbeiten des Vorhabens beschränkt sind, die Daten aber für die in Abschnitt 4 beschriebenen, auch öffentlich zugänglichen lexikographischen Werkzeuge zumindest in Auszügen zur Verfügung stehen.

Frei, d. h. auch ohne Anmeldung zugänglich, sind a) die Texte der „Berliner Zeitung“ (BZ) von 1994 bis 2005. Das BZ- Korpus umfasst 237 Millionen Tokens in 850 000 Dokumenten; b) die online erschienenen Texte des Tagesspiegel (TSP) zwischen 1996 und Juni 2005. Das TSP-Korpus umfasst 156 Millionen Tokens in 330 000 Dokumenten; c) alle Ausgaben der Wochenzeitung *DIE ZEIT*, soweit diese auf zeit.de in digitaler Form zur Verfügung stehen, von 1946 bis heute, sowie Artikel, die nur auf zeit.de online erschienen sind. Das ZEIT-Korpus umfasst 548 Millionen Tokens in 1,2 Millionen Dokumenten. Darüber hinaus können wir, als Ergebnis eines CLARIN-Kurationsprojekts und in Kooperation mit der Staatsbibliothek Berlin und dem Zentrum für Zeithistorische Forschung in Potsdam, die kompletten Jahrgänge der „BZ“ von 1945 bis 1993 (1,6 Millionen Dokumente mit 500 Millionen Textwörtern), des „Neuen Deutschland“ von 1945 bis 1999 (1,45 Millionen Dokumente mit 450 Millionen Tokens) und der „Neuen Zeit“ von 1945 bis 1989 (1,1 Millionen Dokumente mit 400 Millionen Tokens) aus dem „DDR-Presseportal“ zur Recherche für angemeldete Benutzer zur Verfügung stellen.

Eine Reihe weiterer überregionaler Zeitungen, insbesondere die Archivausgaben von WELT, SZ und Bild, wurden ebenfalls in das DWDS aufgenommen. Jedoch stehen nur kleine Teile (3 %) für die Öffentlichkeit zur Verfügung (im Rahmen des Wortprofils, s. Abschnitt 4.2).

Die Zeitungskorpora werden in regelmäßigen Abständen bzw. in dem Maße, wie die Nutzungsverträge mit den Verlagen dies zulassen, aktualisiert.

2.3 Spezialkorpora

Neben den bezüglich der Textsortenauswahl ausgewogenen Kernkorpora und den Zeitungskorpora bieten wir auf unserer Webseite Spezialkorpora an, die hinsichtlich ihres Gegenstandes oder ihrer sprachlichen Charakteristika von den oben genannten Korpora abweichen.

2.3.1 Blog-Korpus

Das *Blog-Korpus* besteht aus Beiträgen und Kommentaren, die in Blogs veröffentlicht worden sind. Diese sind mehrheitlich auf Deutsch und die Betreiber haben die Wiederveröffentlichung der Texte mittels Creative-Commons-Lizenzen ausdrücklich gefördert. Das Korpus wird regelmäßig aktualisiert.Für weitere Details zur Erhebungsmethode vgl. Barbaresi & Würzner (2014). Das Korpus umfasst 102 Millionen Tokens in 231 000 Dokumenten.

2.3.2 Webkorpus

Das nach Anmeldung recherchierbare *Webkorpus* besteht aus einer Auswahl von Webseiten auf Deutsch (vor allem aus Deutschland, Österreich und der Schweiz) und wurde in Zusammenarbeit mit der Österreichischen Akademie der Wissenschaften (Abteilung Academy Corpora) zusammengestellt. Für jede Homepage wurde eine Untermenge von bis zu 500 Seiten heruntergeladen und einer formalen Kontrolle unterzogen: Nur die Dokumente, für die Metadaten (wie das Datum) sowie Text extrahiert werden konnten, wurden in das Korpus aufgenommen. Die Basis besteht aus 200 000 unterschiedlichen Webseiten, die professionell (z. B. Nachrichten- und Firmenseiten) oder privat (u. a. Vereine, Gemeinschaften, Hobbys) betrieben werden. Für weitere Details zum Aufbauverfahren vgl. Barbaresi (2016). Das Korpus umfasst 3 Milliarden Tokens in 6,98 Millionen Dokumenten (Stand Mai 2017). Das Korpus wird fortlaufend aktualisiert.

2.3.3 Untertitelkorpus (1916–2014)

Das *Untertitelkorpus* ist eine Sammlung von Film- und Serienuntertiteln auf Basis des deutschsprachigen Teils der Communityplattform *opensubtitles.org*. Untertitel stellen eine interessante Quelle für Belege gesprochener Sprache dar. Das Untertitelkorpus wurde 2013 erstellt und 2014 aktualisiert. Das Korpus umfasst 75 Millionen Tokens in 12 000 Dokumenten und wird vor allem hinsichtlich der Datenqualität weiter gepflegt. Weitere Details hierzu in Barbaresi (2014).

2.3.4 Dinglers Polytechnisches Journal (1820–1931)

Der große zeitliche Bogen, den Dinglers „Polytechnisches Journal“ mit einer Laufzeit von 111 Jahren (1820–1931) umfasst, spannt sich von den Anfängen der Elektrotechnik bis zum vorläufigen Abschluss der Relativitätstheorie. Waren es zunächst eher Fragen der sich langsam industrialisierenden Agrarkultur und des sich zunehmend mechanisierenden

Handwerks, treten sukzessiv Bereiche wie Bergbau und Hüttenwesen, Maschinen- und Fahrzeugbau, Antriebstechnik, chemische Verfahren, Elektro- oder Nachrichtentechnik hinzu. Das gesamte Werk wurde in einem DFG-Projekt erfasst und vom Deutschen Textarchiv in ein einfach nachnutzbares Format umgewandelt, so dass es für die Recherche im DWDS zur Verfügung steht. Das Korpus umfasst 77 Millionen Tokens in 42 000 Dokumenten[4].

2.3.5 Gesprochene Sprache (1900–2001)

Das Teilkorpus „Gesprochene Sprache“ umfasst Transkripte aus dem gesamten 20. Jahrhundert (Reden, Rundfunkansprachen, Auszüge aus österreichischen Parlamentsprotokollen, Auszüge aus ca. 250 Spiegel-Interviews, Auszüge aus dem *Literarischen Quartett* von 1988 bis 2001, Auszüge aus dem Projekt *Emigrantendeutsch in Israel* sowie Auszüge aus Bundestagsprotokollen von 1998 bis 1999). Das Korpus umfasst knapp 3 Millionen Tokens in 600 Dokumenten.

3. Verwaltung der Korpora

Eine Voraussetzung für die Integration von Korpustexten in das DWDS ist deren strukturelle und linguistische Annotation und die Bereitstellung von Metadaten (Titel, Autorschaft, Erscheinungsdatum u. v. m.). Bei der strukturellen und linguistischen Annotation folgt das DWDS den Empfehlungen der Text Encoding Initiative (TEI)[5], im Speziellen den *Guidelines* in der fünften Version (TEI-P5). Die einzelnen Textwörter werden darüber hinaus mit weiteren, für die linguistische Suche relevanten Informationen versehen. Zur Zeit werden für jedes Textwort die Grundform (Lemma) und die Wortart angegeben und von der Suchmaschine indiziert. Hierfür verwendet das DWDS eine im Hause entwickelte Software, die hier nicht weiter beschrieben werden soll[6].

Alle Korpora des DWDS werden in einem, seit 2014 im Hause entwickelten Korpusverwaltungssystem aufbereitet und gepflegt. Ziel dieses im Hintergrund arbeitenden Systems ist es, einen einheitlichen Zugang zu allen Textkorpora zu ermöglichen, auch und gerade angesichts der zwischen den Korpora bestehenden Unterschiede in den Metadaten. Bereits vorhandene Korpora werden damit auf einen einheitlichen Stand der linguistischen Analyse, der Metadatenindizierung und der Volltextrecherche gebracht, und der Arbeitsaufwand für die Erstellung und Indizierung neu hinzugekommener Korpora wird minimiert. Leitlinien der Entwicklung waren die Festlegung allgemeiner Entwurfsprinzipien, Flexibilität in der Konfiguration und leichte Bedienbarkeit für Korpusverwalter, Entwickler und Endnutzer. Das System umfasst drei Teilsysteme: ein *Bausystem* zur Indizierung neuer Korpora bzw. zur Aktualisierung existierender Korpusindizes, ein *Serversystem* zur Verwaltung von

4Weiterführende Literatur unter http://www.polytechnischesjournal.de/projekt/publikationen/.

5http://www.tei-c.org.

6Details hierzu unter: https://www.dwds.de/d, Abschnitt „Erschließung“.

Korpusserverprozessen und ein *Laufzeitsystem* für den Nutzerzugang.

Das *Bausystem* erlaubt es, aus Rohtextkorpora in einem TEI-Format einen linguistisch annotierten DDC-Suchmaschinenindex (die Suchmaschine wird in Abschnitt 4.1 näher beschrieben) zu bilden. Die händische Auswertung aller indizierten bibliographischen Attribute aus 30 verschiedenen Korpusindizes führte zur Erstellung von „Best Practice"-Richtlinien zur Aufnahme und Benennung bibliographischer Metadaten. Um die Interoperabilität unter den Korpora zu sichern, wurden sinnvolle Defaultwerte durch das Bausystem für alle als obligatorisch eingestuften Metadatenattribute vergeben. Das *Serversystem* ist für die Verwaltung laufender Serverprozesse verantwortlich. Eine webbasierte Oberfläche[7] ermöglicht einen schnellen Überblick über alle Korpora. Insbesondere können dadurch die Anzahl der Subkorpora, der indizierten Dateien und Textwörter sowie das Datum der letzten Indizierung eingesehen werden. Das *Laufzeitsystem* gewährleistet den Zugriff auf die Korpusdaten mittels einer HTTP-basierten RESTful[8]-Schnittstelle zu den Korpusindizes. Optionale korpusspezifische Anpassungen werden zur Laufzeit interpretiert, um eine größtmögliche Flexibilität zu gewährleisten. Für einfache Bedienbarkeit und Interoperabilität sorgen eine umfangreiche Auswahl an Ausgabeformaten: als maschinenlesbare Formate werden z. B. JSON[9], Atom[10], RSS[11] und das in CLARIN-D entwickelte Format TCF[12] angeboten, während menschlichen Nutzern die Suchergebnisse als HTML, KWIC-Zeilen oder als einfacher Text im Browser dargeboten werden.

Da die Korpora des DWDS sowohl Texte mit normierter als auch mit nicht- normierter Orthographie enthalten (letzteres meist bei historischen Texten), wurde mit CAB („Cascaded Analysis Broker") ein Programm zur schreibweisentoleranten linguistischen Analyse historischer Texte entwickelt (Jurish 2013). CAB setzt verschiedene regelbasierte und stochastische Verfahren ein, um historische Schreibvarianten auf äquivalente „kanonische" moderne Wortformen abzubilden. CAB ist in das Bausystem integriert.

Zum Korpusmanagement gehört auch die Umsetzung von Zugriffsbeschränkungen für Teilkorpora, die sich aus Verträgen mit den jeweiligen Lizenzgebern, also den Inhabern von Urheber- und Verwertungsrechten, ergeben. Auf der einen Seite sollte der interessierten Öffentlichkeit so viel wie möglich angezeigt werden können. Zum anderen müssen die Rechte der Lizenzgeber respektiert werden. Um beiden Seiten gerecht zu werden, wurden die folgenden Maßnahmen in das Korpusmanagement integriert: a) für

7 http://kaskade.dwds.de/dstar

8 Fielding, Roy Thomas (2000). "Architectural Styles and the Design of Network-based Software Architectures". *Dissertation*. University of California, Irvine.

9 http://json.org/.

10 https://tools.ietf.org/html/rfc4287.

11 http://www.rssboard.org/rss-specification.

12 http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

jedes Korpus lassen sich die Gruppe der Nutzungsberechtigten und die maximale Länge eines angezeigten Textausschnitts festlegen; b) unter den prinzipiell Nutzungsberechtigten unterscheiden wir angemeldete und nicht-angemeldete Nutzer; c) Teile der Korpora, die nicht als Ergebnisse der Recherche über unsere Suchmaschine angezeigt werden können, finden Eingang in die Datenbanken, auf denen das Wortprofil und der Gute-Belege-Extraktor beruhen (s. unten, Abschnitt 4), und sind über diese Werkzeuge indirekt recherchierbar; d) das Herunterladen von kompletten Texten ist, mit Ausnahme des Deutschen Textarchivs, nicht möglich, auch die Zahl der exportierbaren Belege als Ergebnis einer Suchanfrage ist eingeschränkt.

Das Korpusmanagement des DWDS ist damit auf leichte und effiziente Erweiterbarkeit der Korpusressourcen, auf Interoperabilität zwischen verschiedenen Datenquellen und auf den schnellen möglichst intuitiven Zugriff auf die Daten durch Nutzer mit lexikographischen und allgemeineren linguistischen Fragestellungen ausgerichtet[13].

4. Such- und Analysewerkzeuge

In der modernen, korpusbasierten Lexikographie werden die Aussagen, die Lexikographinnen und Lexikographen zu bestimmen sprachlichen Aspekten und Eigenschaften der beschriebenen Wörter und zu Besonderheiten ihrer Verwendung treffen, auf Korpusevidenz gestützt. Eine Diskussion der Methodik und zahlreiche Beispiele hierfür finden sich in Geyken & Lemnitzer (2016). Allerdings sind der für die Inspektion der Daten verfügbare zeitliche Aufwand und auch die Fähigkeit, bestimmte Muster in den Daten zu erkennen, begrenzt. Korpusbasierte Lexikographie ist deshalb immer auf die Unterstützung durch Software angewiesen. Die Software ist eher und besser in der Lage, große Datenmengen zu aggregieren (für Frequenzen, typische Wortverbindungen und Ähnliches) oder Daten zu extrahieren, die für die lexikographische Beschreibung interessant sind (vgl. hierzu Kilgarriff & Kosem 2012).

Im DWDS-Vorhaben werden diese Vorteile algorithmischer Verfahren der Datenanalyse genutzt, um die Arbeit der Lexikographinnen und Lexikographen zu unterstützen, zu erleichtern und zu beschleunigen.

Fünf Werkzeuge sollen hier erwähnt und im Folgenden kurz beschrieben werden, weil die Resultate dieser Werkzeuge auch den nicht-lexikographischen Nutzern der Webseite zur Verfügung stehen und für sie interessant sein dürften: die linguistische Suchmaschine DDC, zwei Werkzeuge zur Extraktion und Klassifikation von Kollokationen in synchroner (Wortprofil) und diachroner Perspektive (DiaCollo), ein Werkzeug zur Erstellung von Frequenzprofilen von Einzelwörtern über bestimmte Zeitabschnitte (Wortverlaufskurve) und ein Werkzeug zur Ermittlung guter Belege (Gute-Belege-Extraktor).

13 Ein benutzerorientierter Überblick über die Abfragemöglichkeiten ist unter https://www.dwds.de/d/suche abrufbar.

4.1 Die linguistische Suchmaschine DDC

Die im Rahmen des Digitalen Wörterbuchs der deutschen Sprache verwendete linguistische Suchmaschine DDC ist ein speziell auf linguistische und lexikographische Bedürfnisse zugeschnittenes Werkzeug (Jurish et al., 2014). Es verfügt insbesondere über folgende technische Eigenschaften: Boolesche Operatoren; Abstandssuche (sowohl gerichtet als auch symmetrisch); wahlweise satzbasierte oder dokumentbasierte Suche; verschiedene Sortierungsmöglichkeiten, auch nach Metadaten wie dem Erscheinungsdatum eines Textes. In DDC können verschiedene Metadaten indiziert werden. Darüber hinaus werden linguistisch relevante Funktionalitäten angeboten: Wortpositionen können mit beliebig vielen Annotationen versehen sein, nach denen auch gesucht werden kann. Für die DWDS-Suche werden derzeit die Suche nach Wortform, Lemma und Wortart verwendet. Möglich ist die Einbindung von Thesauri. Durch einen in DDC eingebauten Mechanismus kann damit auch nach den Unter- oder Oberbegriffen eines Suchbegriffs gesucht werden. Sowohl die Indizierungs- als auch die Abfragezeiten sind auch für größere Anwendungen hinreichend schnell. Die Abfragezeit für die ersten zehn Treffer bei einfachen Suchabfragen liegt bei ca. 0,05 Sekunden. DDC wird als Webservice auf der Webseite des DWDS eingesetzt. Die dort für die Recherche bereitstehenden Korpora (s. Abschnitt 2) können damit abgefragt und die Ergebnisse in verschiedenen Formaten und nach verschiedenen Kriterien sortiert angezeigt werden.

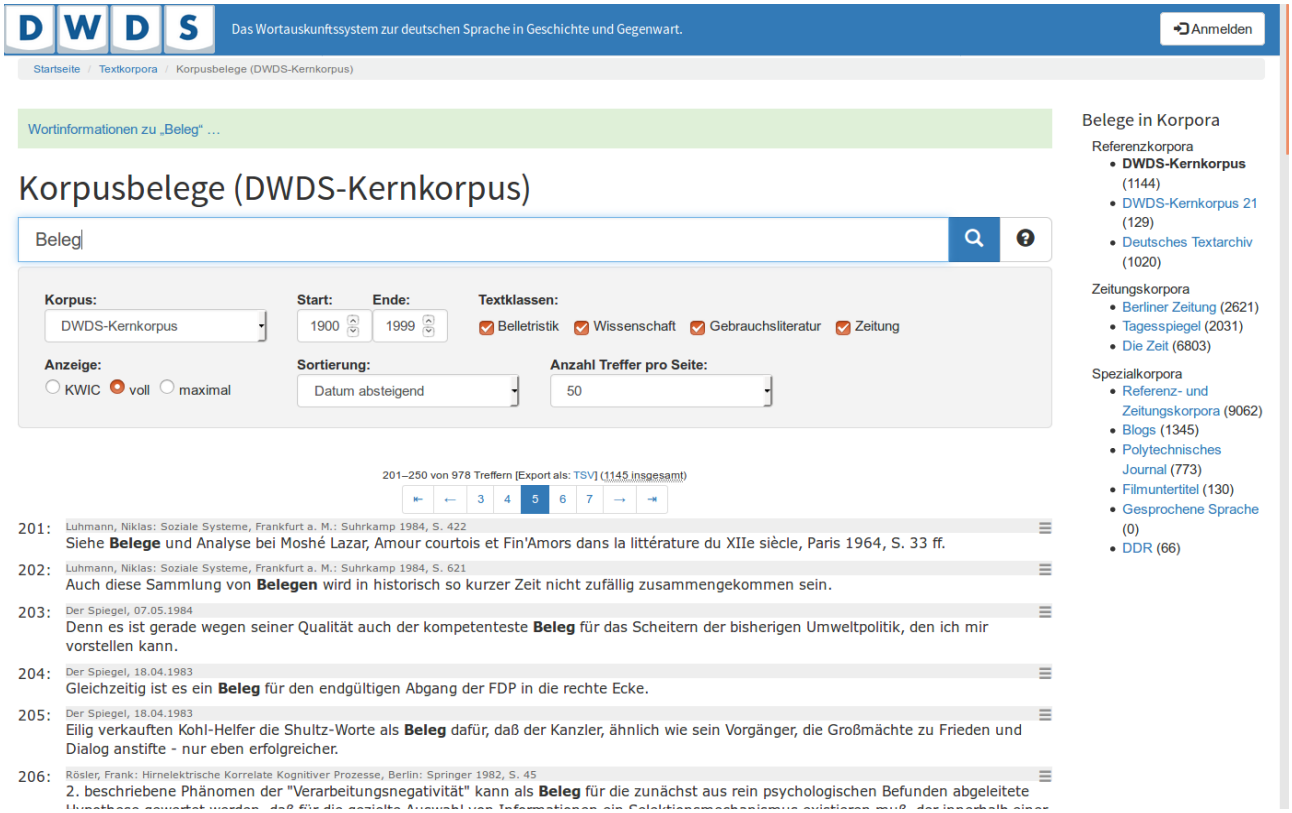

Abb. 1: Suche mit DDC im Kernkorpus nach Vorkommen des Lemmas *Beleg*. Im grau unterlegten Panel lassen sich u.a. das Korpus, der Analysezeitraum, die Beleglänge und die Art der Sortierung auswählen. Angezeigt werden Belege mit ihren Belegquellen.

4.2 Wortprofil

Das DWDS-Wortprofil ist das Ergebnis einer automatischen syntaktischen und statistischen Analyse ausgewählter DWDS-Korpora. Es liefert einen kompakten Überblick über statistisch signifikante und damit typische Wortverbindungen. Beispiele hierfür sind Attribut-Nomen-Verbindungen wie *schöne Bescherung* oder Verb-Objekt-Verbindungen wie *Flasche entkorken*.

Die Darstellung der Wortverbindungen erfolgt in Form einer Schlagwortwolke oder in Tabellenform. Hierbei kann über die einzelnen Verbindungen direkt per Mausklick auf die einzelnen Korpusbelege zugegriffen werden. Des Weiteren ist es auch möglich, zwei Wörter miteinander zu vergleichen. Es können sowohl ihre Gemeinsamkeiten als auch ihre Unterschiede ermittelt werden. Die aktuelle Wortprofil-Version (*Wortprofil 2016*) basiert auf etwa 2,7 Milliarden Tokens und enthält 21,2 Millionen verschiedene Kookkurrenzen für etwa 140 000 Lemmaformen.

Abb. 2: die typischen Verbindungen (Kollokationen) zum Stichwort Beleg, dargestellt als Wortwolke. Im unteren Teil die Belege zur Verbindung *wissenschaftlicher Beleg*.

Das DWDS-Wortprofil ist eine von der *Sketch Engine* (vgl. Kilgarriff et al. 2004) inspirierte Anwendung. Weitere Details hierzu sind in Geyken (2011) und Didakowski & Geyken (2014) zu finden.

4.3 DiaCollo

DiaCollo ist ein Werkzeug für das Auffinden typischer Wortverbindungen (Kollokationen) zu einem Stichwort in einem bestimmten Zeitraum und die visuell aufbereitete Darstellung der Ergebnisse. Die Basis für diese Anwendung sind Korpora mit einer zeitlichen Varianz, insbesondere die Referenzkorpora des DWDS oder die Zeitungskorpora mit einem größeren zeitlichen Umfang. DiaCollo ermittelt statistisch signifikante Kookkurrenzen für ein vorher festzulegendes Messintervall (Jahresschnitte, Dekaden etc.). Da diese Kookkurrenzen für ein Stichwort in jedem Messintervall in der Regel unterschiedlich sind, lässt sich mit Hilfe von DiaCollo die Veränderung der Wortumgebungen und damit auch der Bedeutungen dieses Stichworts über den gesamten Korpuszeitraum nachzeichnen. Wenn es sich bei dem Stichwort um ein Schlüsselwort in politischen oder gesellschaftlichen Diskursen handelt, dann können die Veränderungen in der Verwendung des Wortes auch als Zeichen für politische, kulturelle etc. Veränderungen gedeutet werden. Darüber hinaus kann man mit DiaCollo die parallelen Änderungen in der Verwendungshäufigkeit von Gruppen von Wörtern betrachten und vergleichen. Die Wortgruppen können über die Form oder die Bedeutung der Wörter ausgewählt werden[14].

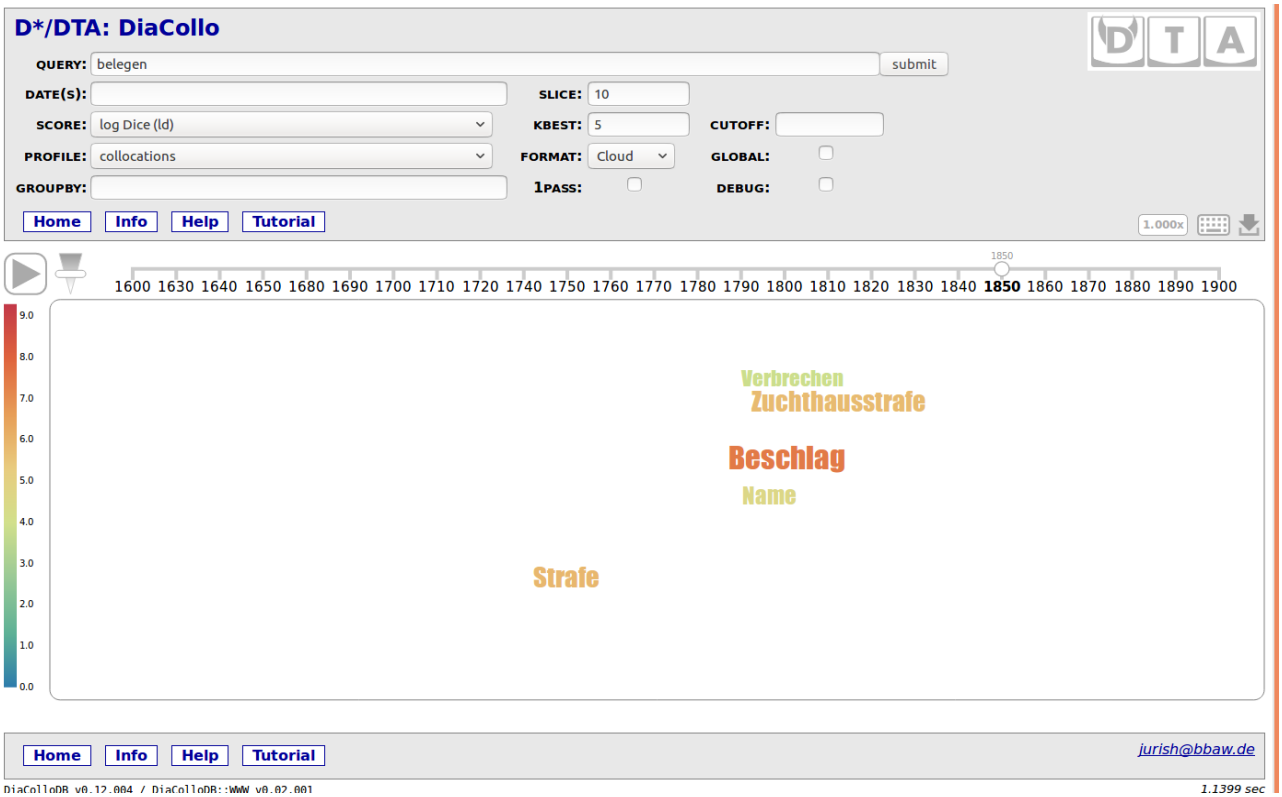

Abb. 3: die fünf typischsten Verbindungen zum Stichwort *belegen* in der Zeitscheibe (Dekade)

14 Ein ausführliches Tutorial mit vielen Beispielen findet ist unter http://kaskade.dwds.de/diacollo-tutorial/ verfügbar, siehe auch Jurish et al. (2016).

um 1850. Als Datenbasis wurde das Kernkorpus des DTA gewählt.

4.4 Wortverlaufskurve

Das DWDS bietet Zugriff auf verschiedene Textkorpora, die insgesamt einen Zeitraum von über 400 Jahren umfassen. Damit ist es möglich, zu jedem Eingabewort ein Profil der Verwendungshäufigkeit eines Wortes innerhalb eines größeren Zeitabschnitts in Form einer Kurve (Verlaufskurve) zu ermitteln. Die Grundlage für die Datenanalyse bilden die in Abschnitt 2 bereits beschriebenen zeitlich und nach Textsortenbereichen ausgewogenen Referenzkorpora: Deutsches Textarchiv (1600–1900), DWDS-Kernkorpus (1900–1999), DWDS-Kernkorpus 21 (2000–2010). Des Weiteren wurden für eine bessere Abdeckung der letzten Dekaden folgende Zeitungskorpora gewählt: Berliner Zeitung (1994–2005), Tagesspiegel (1996–2005), DIE ZEIT (1946–2016). Es werden mehrere Arten der Visualisierung angeboten: normalisiert (d. h. es werden die Frequenzwerte pro Million laufender Textwörter angezeigt) oder in Absolutwerten (d. h. die absoluten Häufigkeiten werden dargestellt). Der Wortverlaufskurve lässt sich z. B. entnehmen, wie sich ein Wort über die Zeit hinweg wandelte („Pressefreiheit“), wann es den Einzug in den Sprachgebrauch schaffte (Beispiel: „Stress“), oder wann es außer Gebrauch kam (Beispiele: „Backfisch“, „apostrophieren“)[15].

Künftig sind zwei Erweiterungen der Wortverlaufskurve vorgesehen. Einerseits soll das Korpus in allen Bereichen, insbesondere des 21. Jahrhunderts substanziell erweitert werden, um für die Gegenwart noch aussagekräftigere Verlaufskurven zu erlangen. Darüber hinaus sollen Wortverlaufskurven auch für den Vergleich von zwei oder mehr Wörtern nutzbar sein.

15 Weitere, auch technische Details zu diesem Service finden sich unter: https://www.dwds.de/d/plot.

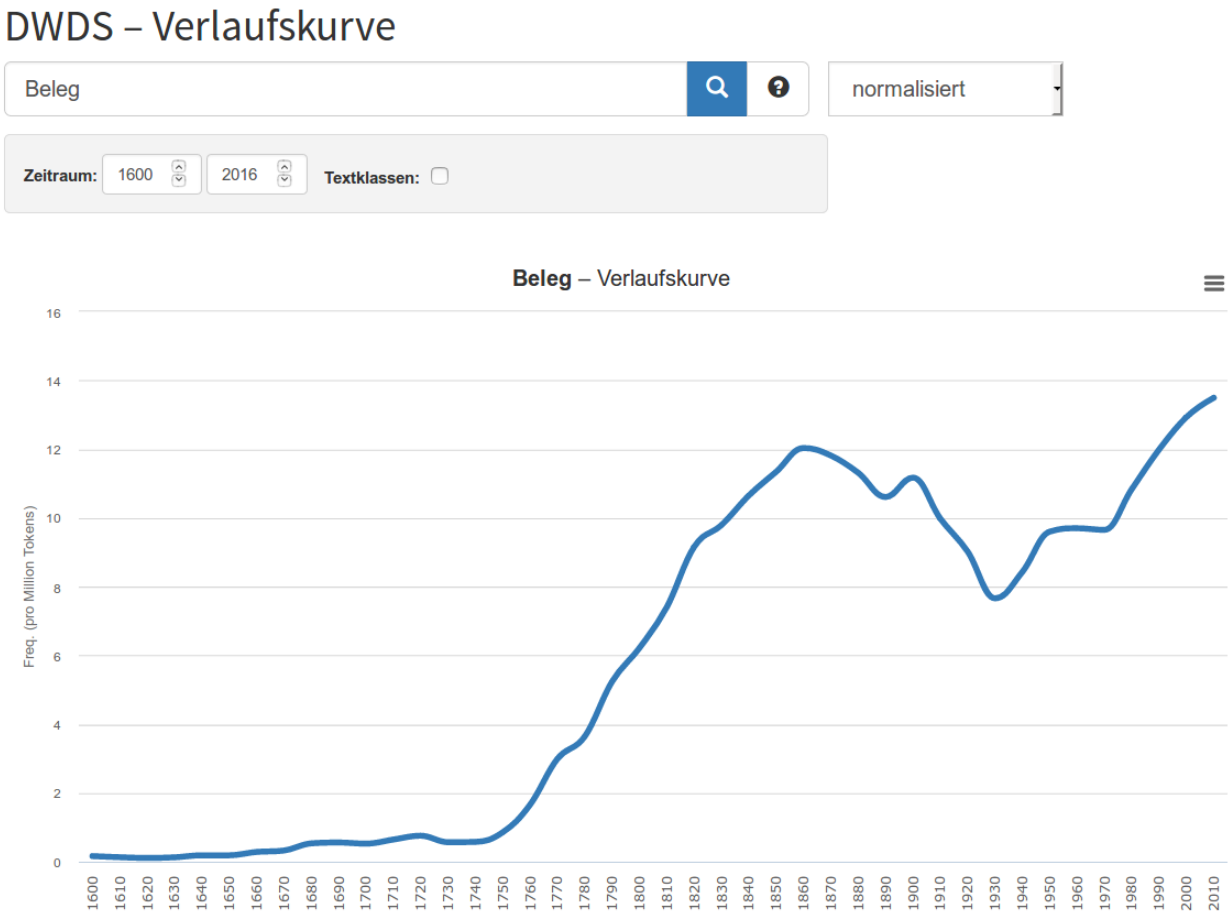

Abb. 4: Wortverflaufskurve für das Stichwort *Beleg*. Die Datenbasis bilden die Kernkorpora (DTA und DWDS).

4.5 Gute-Belege-Extraktor

Der Gute-Belege-Extraktor, dessen Implementierung sich an Kilgarriff et al. (2008) orientiert, soll die lexikographische Arbeit unterstützen und ergänzen. Für die zu bearbeitenden Stichwörter trifft der Extraktor aus einer Vielzahl von Belegen, die für viele Stichwörter in die Tausende oder gar Zehntausende gehen können, eine Vorauswahl nach lexikographischen Kriterien. Die ausgewählten Belege können dann von den belegorientiert arbeitenden Lexikographinnen und Lexikographen mit vertretbarem Aufwand gesichtet werden. Für diejenigen Stichwörter, die aus Zeitgründen nicht mit geprüften Belegen versehen werden können, dient der Extraktor hingegen als Werkzeug zur automatisierten Auswahl und Präsentation guter, also für die Bedeutungserschließung hilfreicher Belege. Aus den beiden genannten Anwendungen ergeben sich konkrete Anforderungen an die Güte der Belege. Sie sollen möglichst selbsterklärend, gut lesbar und verständlich, also kurz, kompakt, syntaktisch nicht zu komplex und typisch für die Verwendung des Stichwortes sein. Das Stichwort sollte im Zentrum des Belegs stehen. Darüber hinaus ergeben sich auch Anforderungen an die Belegzusammenstellung. Diese beinhalten eine gewisse Streuung der Belege über den Beobachtungszeitraum, den Einbezug verschiedener Textsorten und die Bevorzugung „guter" Quellen, wie etwa hochwertiger Zeitungen (eine detaillierte Auflistung dieser Kriterien findet sich in

Didakowski et al. 2012). Diese Anforderungen, vor allem die linguistischen, lassen sich gut operationalisieren und über den Gute-Belege-Extraktor umsetzen, wenn, wie beim DWDS, die Korpustexte über gute Metadaten verfügen und syntaktisch annotiert sind.

Abb. 5: Liste ausgewählter Verwendungsbeispiele für das Stichwort *Beleg*. Es wird ein Belegsatz angezeigt, zusätzlich die Quelle, aus der der Beleg stammt.

Ein Problem, das auch durch das bisher entwickelte Werkzeug nicht befriedigend gelöst werden konnte, ist das Auffinden und die Auswahl passender Belege für seltene Lesarten und Verwendungen des Stichworts im bildlichen oder übertragenen Sinn. Hier ist noch weitere Forschungs- und Entwicklungsarbeit notwendig.

5. Zusammenfassung

In diesem Beitrag haben wir die Korpusplattform des DWDS dargestellt. Die primär für die Zwecke der lexikographischen Arbeit der Projektgruppe erstellten Korpora haben seit der Veröffentlichung der DWDS-Website eine weit über diesen Kreis hinausgehende Nutzung erfahren, insbesondere bei den Nutzerinnen und Nutzern des Wörterbuchs, die die Wörterbucheinträge mit den Textquellen vergleichen wollen, aber auch bei Wissenschaftlerinnen und Wissenschaftlern, die die Korpora des DWDS als Quelle korpuslinguistischer Studien nutzen.

6. Literatur

Adrien Barbaresi. "Efficient construction of metadata-enhanced web corpora." Proceedings of the 10th Web as Corpus Workshop, ACL, 2016, p. 7-16.

Adrien Barbaresi, 2014: *Language-classified Open Subtitles (LACLOS): Download, extraction, and quality assessment.* [Research Report] BBAW. 2014.

Adrien Barbaresi, Kay-Michael Würzner, 2014: *For a fistful of blogs: Discovery and comparative benchmarking of republishable German content.* In: Proceedings of NLP4CMC workshop (KONVENS 2014), Hildesheim University Press, S. 2-10.

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, Angelika Storrer, 2013: DeRik – A German Reference Corpus of Computer-Mediated Communication. In: Literary and Linguistic Computing, 28(2013)4, pp. 531-537

Jörg Didakowski, Alexander Geyken, Lothar Lemnitzer, 2012: Automatic example sentence extraction for a contemporary German dictionary. In: Proceedings EURALEX 2012, Oslo, S. 343-349.

Jörg Didakowski, Alexander Geyken (2014): *From DWDS corpora to a German word profile–methodological problems and solutions.* In: OPAL – Online publizierte Arbeiten zur Linguistik 2/2014, S. 39–47.

Alexander Geyken, 2007: *The DWDS corpus: A reference corpus for the German language of the 20th century.* In: Fellbaum, Christiane (Hg.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects.* London, S. 23–41.

Alexander Geyken, 2011: *Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora.* In: A. Abel & R. Zanin (Eds.), *Korpora in Lehre und Forschung* (pp. 115–137). Bozen, Italien: Bozen University Press.

Alexander Geyken, Lothar Lemnitzer, 2012: Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In: Proceedings EURALEX 2012, Oslo, S. 362-366.

Alexander Geyken, Lothar Lemnitzer, 2016: Automatische Gewinnung von lexikographischen Angaben. In: DfG-Netzwerk Internetlexikografie unter Leitung von Annette Klosa und Carolin Müller-Spitzer: Internetlexikografie. Ein Kompendium. Berlin/Boston:de Gruyter 2016, S. 197-247.

Bryan Jurish, 2013: Canonicalizing the deutsches Textarchiv. In I. Hafemann (Hg.): Perspektiven einer corpusbasierten historischen Linguistik und Philologie (Berlin 12.-13. Dezember 2011), Berlin-Brandenburgische Akademie der Wissenschaften.

Bryan Jurish, Christian Thomas, Frank Wiegand, 2014: Querying the Deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Hg.), Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities, S. 25-30.

Bryan Jurish, Alexander Geyken, Thomas Werneke, 2016: DiaCollo – diachronen Kollokationen auf der Spur. In DHd 2016: Modellierung – Vernetzung – Visualisierung, Leipzig, 7.-12. März, S. 172-175.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, David Tugwell, 2004: The sketch engine. In *Proceedings EURALEX 2004*, Lorient, France; Pp. 105–116.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý, 2008: GDEX: Automatically Finding Good Dictionary Examples in a Corpus, In Bernal, Elisenda / DeCesaris, Janet (Eds.), Proceedings

of the Thirteenth EURALEX International Congress, Barcelona, Spain, pp. 425-432.

Adam Kilgarriff, Iztok Kosem, 2012: Corpus Tools for Lexicographers. In: *Electronic Lexicography* Sylviane, ed. By Sylviaine Granger and Magali Paquot (eds.) Oxford Univ Press, pp. 31-55.

Wolfgang Klein, Alexander Geyken, 2010: Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich/Schierholz, Stefan/Schweickard, Wolfgang/Wiegand, Herbert Ernst/Gouws, Rufus H./Wolski, Werner:Lexikographica. Berlin/New York, S. 79-93.

Alexey Sokirko, 2003. DDC – A search engine for linguistically annotated corpora. In: Proceedings of Dialogue 2003, Protvino, Russia.