Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie

Angelika Storrer

Preprint. Erscheint in: Karlfried Knapp u.a. (Hrsg.): Angewandte Linguistik. Ein Lehrbuch. 3. Auflage. Tübingen: Francke Verlag.

1 Korpuslinguistik und angewandte Linguistik

Korpora wurden schon vor der Verbreitung digitaler Medien in vielen Anwendungsfeldern der Linguistik genutzt. Insbesondere in der historischen Sprachwissenschaft und der Lexikographie hat das Sammeln und Auswerten von Belegen aus einem Korpus von Quellentexten eine lange Tradition. Auch in Gesprächforschung und Konversationsanalyse hat man schon vor der Verbreitung digitaler Korpustechnik mit transkribierten Gesprächskorpora gearbeitet. Die computertechnische Speicherung und Auswertung von Korpusdaten bietet nun viele neue Möglichkeiten, sprachliche Regularitäten und Strukturen in authentischen Verwendungskontexten qualitativ und quantitativ zu analysieren. Die dafür relevanten Konzepte und Methoden stammen überwiegend aus der Korpuslinguistik, einem derzeit sehr aktiven Forschungsfeld, in dem Informatik, Computerlinguistik und Linguistik interdisziplinär zusammenarbeiten, um Standards und Werkzeuge für die digitale Erschließung von Korpora zu entwickeln, die als empirische Basis für die Theoriebildung und die Überprüfung theoretischer Annahmen an authentischen Sprachdaten genutzt werden können. Zur Korpuslinguistik gibt es inzwischen sehr gute Einführungen und Überblicksdarstellungen, auf die ich in den einzelnen Kapiteln verweisen werde. Zwei empfehlenswerte Einführungen in die Anwendungsfelder digitaler Korpora sind Lemnitzer/Zinsmeister (2006; Schwerpunkt deutsche Sprache) und McEnery/Xiao/Tono (2006; auf Englisch), beide sind verständlich und anwendungsbezogen geschrieben, beide diskutieren den Einsatz von Korpora in verschiedenen Anwendungsfeldern am Beispiel publizierter korpusgestützter Fallstudien. Für die vertiefende Lektüre zu speziellen Aspekten empfehlen sich die Artikel der beiden aktuellen HSK-Handbücher zum Thema (Lüdeling/Kytö 2008/2009).

Aus den Anwendungsfeldern der Korpuslinguistik greife ich in diesem Artikel die empirische Erforschung der Wortschatzentwicklung in der Lexikographie und der Phraseologie heraus. In beiden Bereichen werden digitale Korpora inzwischen intensiv genutzt: Die meisten aktuelle Wörterbuchprojekte arbeiten mit digitalen Korpora; für die Beschreibung von Mehrwortlexemen und Kollokationen existieren spezialisierte korpusbasierte Werkzeuge. Im Internet entstehen digitale lexikalische Informationssysteme, in denen Wörterbücher, Korpora und Korpusauswertungswerkzeuge unter einer einheitlichen Nutzeroberfläche angeboten werden. Nutzer derartiger Systeme können Eigen-

schaften einer Wortschatzeinheit nicht nur in den Wörterbuchartikeln nachschlagen, sondern auch eigene Recherchen anstellen, z. B. um nach typischen Verwendungskontexten oder ungewöhnlichen Verwendungsweisen zu suchen oder um Prozesse der Bedeutungsveränderung über einen bestimmten Zeitraum hinweg nachzuverfolgen. Sprachinteressierte und „Spracharbeiter" in Verlagen und Bildungsinstitutionen verfügen damit online und kostenfrei über Möglichkeiten zur eigenständigen Sprachanalyse, die bislang den Wörterbuchredaktionen vorbehalten waren. Ziel des Artikels ist es, das Hintergrundwissen einzuführen, das man für die Nutzung derartiger Systeme benötigt, und die Potenziale der korpusgestützten Sprachanalyse an einfachen Analysebeispielen zu illustrieren. Die Beispiele werden ergänzt durch Verweise auf Literatur zu weiterführenden methodischen und korpuslinguistischen Fragen.

Der Artikel ist folgendermaßen aufgebaut. Im nächsten Kapitel werden diejenigen Grundbegriffe digitaler Korpustechnologie eingeführt, die für die Auswahl eines zu einer Fragestellung passenden Korpus sowie für das Verständnis von Meta- und Hilfetexten der Online-Korpora unabdingbar sind. Abschnitt 3 gibt einen Überblick über wichtige Korpusressourcen für das Deutsche. In Abschnitt 4 werden die Einsatzmöglichkeiten linguistisch annotierter Korpora für die lexikographische Sprachanalyse an Beispielen erläutert.

2 Grundbegriffe der korpusgestützten Sprachanalyse

In ihrer Einführung in die Korpuslinguistik definieren Lemnitzer/Zinsmeister (2006:7) den Ausdruck ‚Korpus' wie folgt: „Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d. h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind." Die in dieser Definition enthaltenen Begriffe ‚Daten', ‚Metadaten' und ‚Annotationen' werden in Abschnitt 2.1. erläutert. Der Zugriff auf digital gespeicherte Korpora erfolgt über Nutzerschnittstellen, die man als Korpusrecherchesysteme bezeichnet. Die zentralen Funktionen solcher Systeme sind in Abschnitt 2.2. beschrieben. Das Vorhandensein von Metadaten, linguistischen Annotationen und spezialisierter Recherchesoftware unterscheidet linguistisch aufbereitete Textkorpora von digitalisierten Textsammlungen und von Suchwerkzeugen im World Wide Web. Die Eigenschaften linguistisch aufbereiteter Textkorpora und ihr Verhältnis zu anderen digitalen Datensammlungen werden in Abschnitt 2.3. erläutert.

Beim Sprechen und Schreiben über Korpora ist es hilfreich, den alltagssprachlichen Ausdruck ‚Wort' zu präzisieren. In diesem Artikel geschieht dies mit Hilfe der Termini ‚Wortvorkommen', ‚Wortform' und ‚Lexem', die in folgendem Verhältnis zueinander stehen:

(1) Als ‚Wortvorkommen' zählt jedes Vorkommen eines Wortes in einem fortlaufenden Text. Als alternative Bezeichnungen für diesen Typ von Einheit findet man auch ‚(das) Token','Textwort' oder ‚laufendes Wort'. Wenn man den Beispielsatz *to be or not to be that is the question* segmentiert, erhält man also zehn Wortvorkommen.

(2) Die Einheit ‚Wortform' ist über ihre Form bestimmt, unabhängig davon, wie häufig diese in einem Satz oder Text vorkommt. Der o. g. Beispielsatz enthält demnach acht Wortformen.

(3) Für die semantisch bestimmten Wortschatzeinheiten, die im Regelfall Gegenstand lexikologischer und lexikographischer Analysen sind, verwende ich den Terminus ‚Lexem'. Im Kontext der Lexikographie sind auch die Bezeichnungen ‚Lemma' bzw. ‚Stichwort' gebräuchlich. In flektierenden Sprachen bilden Lexeme bestimmter Wortklassen mehrere Wortformen aus; im o. g. Beispielsatz würde man z. B. die Wortformen *is* und *be* demselben Lexem zurechnen; der Satz enthält also sieben Lexeme.

Der Umfang von Korpora wird meist in Wortvorkommen bemessen; bei manchen Korpora wird zusätzlich die Zahl der Wortformen, der Sätze oder der Dokumente (Texte bzw. Gesprächsmitschnitte) angegeben. Die Suche in Korpora operiert vornehmlich auf Wortformen und nicht auf Lexemen. Wer in einem Korpus nach einem bestimmten Lexem sucht, wird damit rechnen müssen, auch homographe Wortformen anderer Lexeme in der Trefferliste zu finden. Denn ohne linguistische Annotationen ist es beispielsweise nicht möglich, zwischen der Wortform *ein* als unbestimmtem Artikel (*ein Auto*) und als abtrennbarem Verbzusatz (*sie finden sich dort ein*) zu differenzieren.

Bei der Beschreibung von lexikographischen Korpusrecherchen ist es wegen dieses Homographieproblems sehr hilfreich, terminologisch zu differenzieren zwischen (1) der ‚Trefferliste', die vom Korpusrecherchesystem automatisch zu einer Suchanfrage erstellt wird, und (2) der ‚Belegliste', dem Resultat einer intellektuellen Nachbearbeitung. Insgesamt verwende ich in diesem Artikel den Ausdruck ‚Treffer' zur Bezeichnung der Korpussegmente, die ein Korpusrecherchesystem als passend für eine Suchanfrage ausgibt. Mit dem Ausdruck ‚Beleg' bezeichne ich die Teilmenge der Treffer, die für mein Untersuchungsziel auch tatsächlich relevant sind. Treffer in der Trefferliste, die für mein Untersuchungsziel nicht relevant sind, bezeichne ich als ‚Pseudotreffer'. Mit diesen terminologischen Vereinbarungen lässt sich ein typischer Arbeitsablauf bei der korpusgestützten Sprachanalyse folgendermaßen beschreiben: Man formuliert für eine Untersuchungsfrage ein Suchmuster; das Korpusrecherchesystem generiert dazu eine Trefferliste. In der Abfrage in Abb. 3 sollten Belege für das Verb *einfinden* gesucht werden; die Treffer 2 und 3 erweisen sich für diese Abfrage als Pseudotreffer. Wenn man derartige Pseudotreffer aus der Trefferliste entfernt, erhält man eine Belegliste, die dann nach weiteren Gesichtspunkten geordnet und weiter bearbeitet werden kann.

2.1 Primärdaten – Metadaten – Annotationen

Die in digitalen Korpora gespeicherten Daten (Textdokumente, Gesprächstranskriptionen, Bild-, Ton und Videodateien) bezeichnet man als Primärdaten, wenn es darum geht, sie von den Metadaten abzugrenzen, also von Daten, mit denen die Primärdaten näher beschrieben und klassifiziert sind. Typische Metadaten zu Korpora geschriebener Sprache sind Autor, Erscheinungsdatum und Publikationsort. Typische Metadaten zu Gesprächskorpora sind Aufnahmedatum, -ort, und -dauer, Informationen zu den Gesprächsbeteiligten und zum Thema/Anlass der Interaktion sowie ggf. Angaben zum Transkriptionsstandard. Zu den Metadaten zählt man auch die Zuordnung zu Sprachen, Text- bzw. Gesprächssorten oder Themengebieten bzw. Rubriken (in Zeitungskorpora). In der Lexikographie sind Metadaten mit exakten Quellenangaben (Autor, Publikationsort mit Seitenangabe) wichtig, um Belege zitierbar zu machen. Korpusrecherchesysteme können digital verwaltete Metadaten nutzen, um Suchanfragen auf bestimmte Autoren, Zeitspannen oder Textsortenbereiche einzuschränken; Zeitungskorpora bieten oft Suchfilter nach Rubriken und Themen. Weiterhin können Metadaten in die automatische Auswertung der Primärdaten einfließen. Das automatisch generierte Frequenzverlaufsdiagramm in Abb. 5 wird beispielsweise auf der Basis von Metadaten zum Erscheinungsjahr und zum Textsortenbereich der Texte aus dem Kernkorpus des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS; s. Abschnitt 2.3) generiert.

Als linguistische Annotationen bezeichnet man Informationen zu linguistischen Merkmalen, die den Primärdaten des Korpus in digitaler Form beigefügt sind. Wie schon die Metadaten dienen auch die Annotationen primär dazu, die Suchpräzision und die automatische quantitative Auswertung der Korpusdaten zu verbessern. Einen Überblick über Verfahren und Nutzwert linguistischer Annotationen, der auch für computerlinguistische Laien gut verständlich ist, bieten Lemnitzer/Zinsmeister (2006:Kap. 4). Ich möchte im Folgenden nur die Grundbegriffe herausgreifen, die für das Verständnis der in diesem Artikel besprochenen Analysebeispiele relevant sind.

Ein wichtiger Typ von linguistischer Aufbereitung ist die Wortartenannotation (engl. part-of-speech tagging, POS-Tagging). Sie dient primär dazu, homographe Wortformen – z. B. *schicke* als Wortform des Adjektivs *schick* und *schicke* als Imperativform des Verbs *schicken* – zu vereindeutigen. Bei der Wortartenannotation wird jedem Wortvorkommen im Korpus ein Kürzel zugeordnet, das für eine syntaktische Kategorie steht. Die Kürzel bezeichnet man in der Korpuslinguistik als ‚Tags'; das Inventar und die Bedeutung der Tags wird in ‚Tagsets' festgelegt. Ein für die deutsche Korpustechnologie bedeutsames standardisiertes Tagset ist das Stuttgart-Tübingen-Tagset STTS, das Tags für die Wortartenannotation deutscher Korpora bereitstellt (dokumentiert in [STTS], einführend: Lemnitzer/Zinsmeister 2006:68f.). Den Wortvorkommen des Satzes *Peter hält an der Ampel an.* würden nach diesem Standard die folgenden Tags zugeordnet (in Spitzklammern sind die Kategorien der Tags erläutert):

Peter: NE <Eigennamen>
hält: VVFIN <finites Verb, voll>
an: APPR <Präposition>
der: ART <bestimmter bzw. unbestimmter Artikel>
Ampel: NN <„normales“ Nomen>
an: PTKVZ <abtrennbarer Verbzusatz>

Auf der Basis derartiger Annotationen kann man beispielsweise Eigennamen (Tag: NE) von Gattungsbezeichnungen (Tag: NN) unterscheiden – bei der lexikographischen Bearbeitung von Stichwörtern wie *Hahn*, *Schlauch* oder *Fischer* kann allein dadurch die Zahl der Pseudotreffer in einer Trefferliste erheblich reduziert werden. Das Tag PTKVZ für abtrennbare Verbzusätze erleichtert die Suche nach deutschen Partikelverben (wie z. B. *einfinden, anhalten, zumachen*), denn zu sehr vielen trennbaren Verbpartikeln existieren hochfrequente homographe Formen, die anderen syntaktischen Kategorien angehören (vgl. hierzu die Trefferlisten in den Abb. 2 und 3)

Eine weitere wichtige Form der linguistischen Aufbereitung für eine flektierende Sprache wie das Deutsche ist die Lemmatisierung, bei der flektierte Wortformen (*sah*, *sieht*, *sähe*, *gesehen*) auf eine Grundform (das Lemma *sehen*) zurückgeführt werden. Auf der Basis kann ein Korpusrecherchesystem nicht nur eine wortformbasierte Suche anbieten, sondern auch eine lemmabasierte Suche, bei der alle flektierten Formen zur Grundform ausgegeben werden. Interessant für wortgrammatische Analysen ist auch die morphologische Analyse, die es ermöglicht, gezielt nach bestimmten Wortstrukturen zu suchen, z. B. nach Komposita mit dem Erstglied *Bären-* (*Bärenhunger*, *Bärendienst*) oder Ableitungen mit dem Präfix *re-* (*reanimieren*, *redistribuieren*). Zwar bieten viele Recherchesysteme in ihrer Abfragesprache Platzhaltersymbole für beliebige Graphemfolgen an. Wer aber in einem morphologisch nicht weiter annotierten Korpus mit dem Suchmuster „re*" nach Belegen für das Präfix *re-* sucht, erhält mehr Pseudotreffer (*reden*, *reisen*, *rennen* etc.) als Belege. Für solche Analysen wäre eine morphologische Aufbereitung sehr hilfreich; bislang wird sie aber noch von keinem der in Abschnitt3 beschriebenen großen deutschen Online-Korpora angeboten.

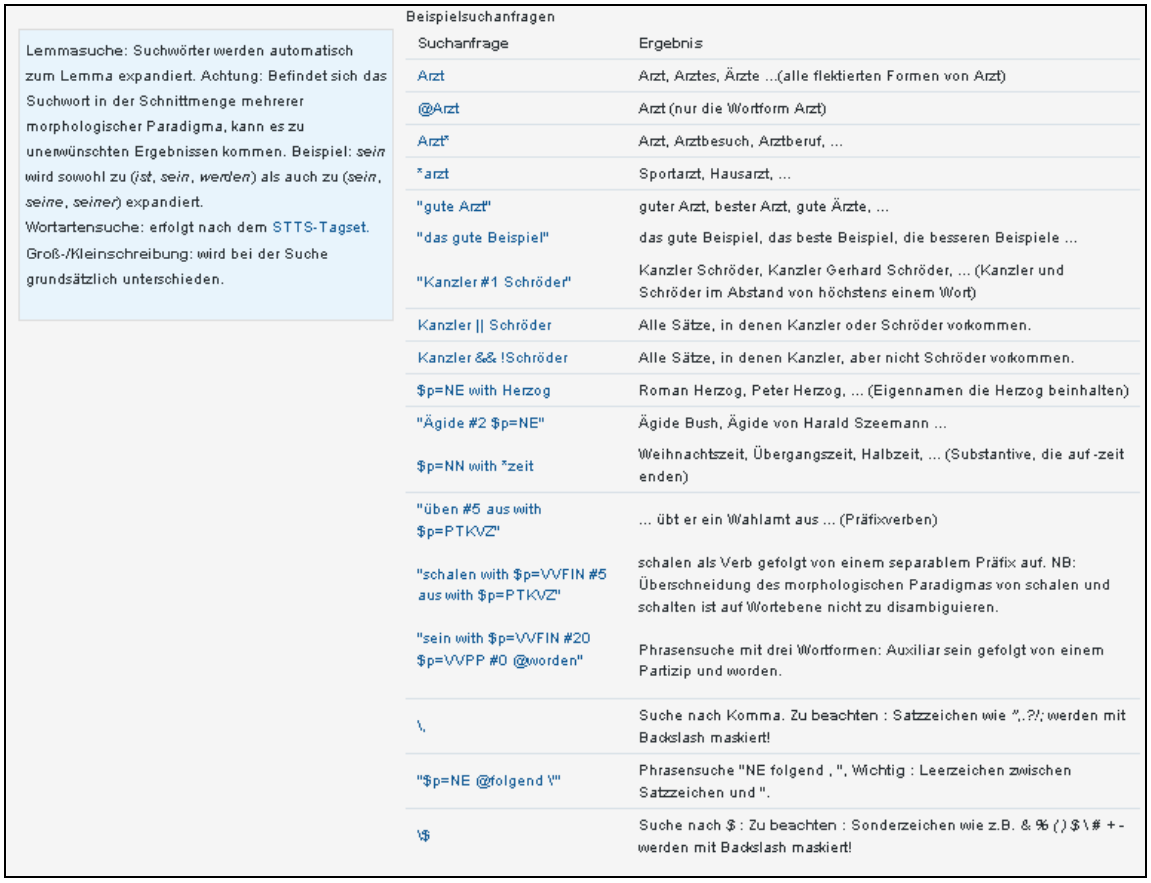
Wortartenannotation und Lemmatisierung gehören zu den Standards der linguistischen Aufbereitung. In großen Korpora erfolgt die Aufbereitung allerdings nicht intellektuell, sondern automatisch – schließlich wäre es extrem aufwändig, 100 Millionen Wortvorkommen oder mehr manuell zu annotieren. Bei der automatischen Wortartenannotation werden meist regelbasierte und statistische Verfahren kombiniert; für das Deutsche gibt es verschiedene, gut entwickelte Werkzeuge (vgl.: Lemnitzer/Zinsmeister 2006:71ff.). Fehlerfreie automatische Zuordnungen kann man allerdings auch von guten Systemen nicht erwarten.

Auch die automatische Lemmatisierung funktioniert in keinem verfügbaren System fehlerfrei; problematisch sind vor allem Lexeme, die teilweise homographe Flexionsformen ausbilden, also z. B. die Verben *fahren* und *führen* oder das Verb *zeitigen* und das Adjektiv *zeitig*. Deshalb erlebt man bei der lemmabasierten Suche immer wieder Überraschungen: Wer denkt schon daran, dass die Wortform *heute* auch eine Flexionsform des Verbs *heuen* (= Heu ernten) ist oder dass die Wortform *weil* auch als Imperativform des Verbs *weilen* interpretiert werden kann. In jedem Fall muss man auch bei einer lemmabasierten Suche mit Pseudotreffern rechnen und auch bei der Interpretation von automatisch erstellten Frequenzangaben sollte man derartige Überschneidungen mit bedenken.

In der Korpuslinguistik wird an Verfahren und Standards zur Annotation syntaktischer Strukturen gearbeitet. Man unterscheidet zwischen syntaktisch partiell annotierten Korpora und syntaktisch vollständig annotierten Korpora (vgl. den Überblick in Lemnitzer/Zinsmeister 2006:74ff.). In syntaktisch partiell annotierten Korpora werden Folgen von Wortvorkommen als Phrasen eines bestimmten Typs annotiert. In den automatisch erzeugten Wortprofilen des DWDS-Korpus kann man aufbauend auf eine derartige Annotation nach Kollokationspartnern eines bestimmten Typs suchen, z. B. nach typischen Akkusativobjekten zum Verb *zeitigen* (vgl. Abb. 6 in Abschnitt 4.4). Für die gezielte Recherche nach syntaktischen Konstruktionen eines bestimmten Typs eignen sich vollständig syntaktisch annotierte Korpora, so genannte Baumbanken (engl. tree banks). Die zur syntaktischen Annotation verwendeten Kategorien variieren in Abhängigkeit vom zugrunde liegenden Grammatikmodell. Eine gut verständliche Einführung in die grundlegenden Konzepte, die für die Analyse von Baumbanken benötigt werden, geben Lemnitzer/Zinsmeister (2006:80ff).

In vielen Kontexten der korpusgestützten lexikographischen Analyse würde man gerne gezielt Belege für eine bestimmte semantische Lesart eines Lexems suchen können, z. B. Belege für *Ampel* als ‚Hängelampe' (in Abgrenzung zu *Ampel* als ‚Verkehrssignal' oder als Kurzwort für *Ampelkoalition*, vgl. das Beispiel in Abschnitt 4.3). Es wäre auch wünschenswert, in einem Frequenzverlaufsdiagramm wie dem in Abb. 5 gezeigten nicht nur nach Textsortenbereichen, sondern auch nach semantischen Lesarten zu differenzieren. Für derartige Funktionen benötigte man jedoch eine semantische Annotation, die jedes Wortvorkommen im Korpus einer semantischen Lesart zuordnet. Leider gehört aber die automatische Disambiguierung von Lesarten im Kontext (engl. word sense disambiguation WSD) trotz langjähriger Forschung immer noch zu den noch nicht befriedigend gelösten Aufgaben der Sprachtechnologie (vgl. den Überblick in Rayson/Stevenson 2008). Eine manuelle Annotation wäre bei großen Korpora zu aufwändig. Man kann also derzeit und ggf. auch noch in absehbarer Zukunft große Korpora nicht automatisch nach disambiguierten semantischen Lesarten durchsuchen. Diese „semantische Blindheit" der aktuellen Korpustechnologie erfordert in vielen Fällen ma-

nuelle Nachbearbeitung – gerade Einsteiger in die korpusgestützte Sprachanalyse sind hierüber oft enttäuscht. Wer häufiger mit Korpora arbeitet, wird allerdings bald ein Gespür für den Zeitaufwand und die richtigen Analysestrategien entwickeln. Dennoch muss das Problem der semantischen Blindheit gerade bei der Bewertung statistischer Ergebnisse im Auge behalten werden, denn auch die Statistiken operieren nicht über Bedeutungseinheiten, sondern über Formeinheiten (vgl. das Beispiel *Ampel* in Abschnitt 4.3).



Lemmasuche: Suchwörter werden automatisch zum Lemma expandiert. Achtung: Befindet sich das Suchwort in der Schnittmenge mehrerer morphologischer Paradigma, kann es zu unerwünschten Ergebnissen kommen. Beispiel: *sein* wird sowohl zu (*ist*, *sein*, *werden*) als auch zu (*sein*, *seine*, *seiner*) expandiert.
Wortartensuche: erfolgt nach dem STTS-Tagset.
Groß-/Kleinschreibung: wird bei der Suche grundsätzlich unterschieden.

Beispielsuchanfragen

Suchanfrage	Ergebnis
Arzt	Arzt, Arztes, Ärzte ...(alle flektierten Formen von Arzt)
@Arzt	Arzt (nur die Wortform Arzt)
Arzt*	Arzt, Arztbesuch, Arztberuf, ...
*arzt	Sportarzt, Hausarzt, ...
"gute Arzt"	guter Arzt, bester Arzt, gute Ärzte, ...
"das gute Beispiel"	das gute Beispiel, das beste Beispiel, die besseren Beispiele ...
"Kanzler #1 Schröder"	Kanzler Schröder, Kanzler Gerhard Schröder, ... (Kanzler und Schröder im Abstand von höchstens einem Wort)
Kanzler || Schröder	Alle Sätze, in denen Kanzler oder Schröder vorkommen.
Kanzler && !Schröder	Alle Sätze, in denen Kanzler, aber nicht Schröder vorkommen.
$p=NE with Herzog	Roman Herzog, Peter Herzog, ... (Eigennamen die Herzog beinhalten)
"Ägide #2 $p=NE"	Ägide Bush, Ägide von Harald Szeemann ...
$p=NN with *zeit	Weihnachtszeit, Übergangszeit, Halbzeit, ... (Substantive, die auf -zeit enden)
"üben #5 aus with $p=PTKVZ"	... übt er ein Wahlamt aus ... (Präfixverben)
"schalen with $p=VVFIN #5 aus with $p=PTKVZ"	schalen als Verb gefolgt von einem separablem Präfix auf. NB: Überschneidung des morphologischen Paradigmas von schalen und schalten ist auf Wortebene nicht zu disambiguieren.
"sein with $p=VVFIN #20 $p=VVPP #0 @worden"	Phrasensuche mit drei Wortformen: Auxiliar sein gefolgt von einem Partizip und worden.
\,	Suche nach Komma. Zu beachten : Satzzeichen wie *".,?!;* werden mit Backslash maskiert!
"$p=NE @folgend \'"	Phrasensuche "NE folgend , ", Wichtig : Leerzeichen zwischen Satzzeichen und ".
\$	Suche nach $: Zu beachten : Sonderzeichen wie z.B. & % ( ) $ \ # + - werden mit Backslash maskiert!

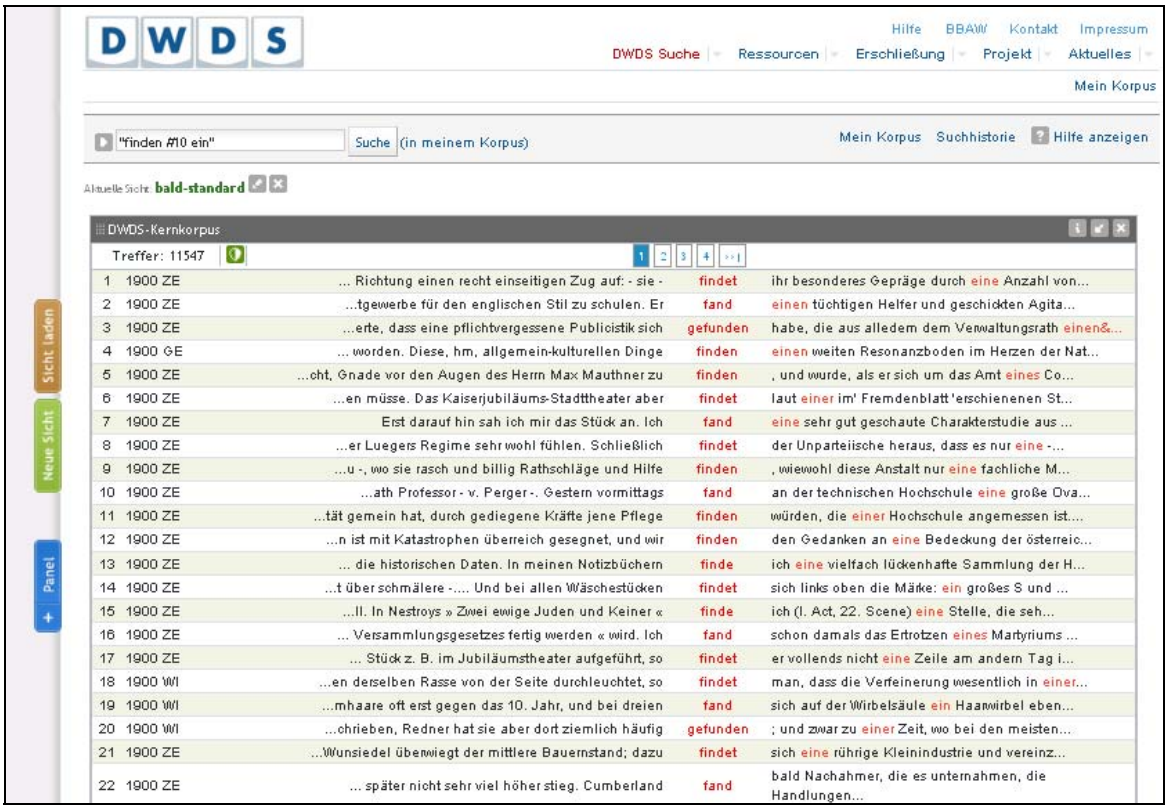
Abb. 1: Syntax für Suchanfragen im DWDS-Korpusrecherchesystem (www.dwds.de)

2.2 Korpusrecherche: Werkzeuge und Funktionen

Um von linguistisch aufbereiteten Korpora profitieren zu können, benötigt man ein Korpusrecherchesystem, das Daten, Metadaten und Annotationen in linguistisch aufbereiteten digitalen Korpora sucht, anzeigt und quantitativ auswertet. Die in Abschnitt 3 genannten Online-Korpora für das Deutsche verfügen über integrierte Recherchesysteme, die man mit einem Webbrowser direkt nutzen kann. Um erste Erfahrungen mit korpusbasierten Analysen zu machen, ist die Nutzung von Online-Korpora mit integrierten Recherchesystemen der schnellste und einfachste Weg. Wer mit selbst zusammengestellten Korpora arbeiten möchte oder muss, findet inzwischen auch hierfür eine Reihe

von kostenfrei verfügbaren Werkzeugen (vgl. die Übersicht in Lemnitzer/Zinsmeister 2006:88ff); ein wenig mehr Zeit und technisches Know-how muss man bei der Arbeit mit eigenen Korpora dennoch mitbringen. Die in Korpusrecherchesystemen angebotenen Funktionen sind ähnlich und in den zugehörigen Hilfetexten im Web auch ausführlich dokumentiert; ich beziehe mich im Folgenden auf die Werkzeuge und Funktionen des DWDS-Systems und die in Abschnitt 4 diskutierten Analysebeispiele.

Ein Korpusrecherchesystem interpretiert Suchanfragen, generiert dazu Treffermengen und zeigt diese an. Die Suchanfragen müssen in bestimmter Form (der Syntax der Abfragesprache) formuliert werden; in Abb. 1 sind die wichtigsten Elemente der Syntax der DWDS-Abfragesprache an Beispielen erläutert. Einige Funktionen kennt man aus der Nutzung von Suchmaschinen im WWW, wenngleich dafür teilweise andere Symbole verwendet werden: In der Syntax des DWDS-.Systems sucht man mit „a && b“ nach dem gemeinsamen Vorkommen der Suchwörter a und b im Satz (Und-Verknüpfung). Zur Anfrage „a || b“ passen alle Sätze, in denen entweder das Wort a oder das Wort b vorkommt (Oder-Verknüpfung). Nach einer speziellen Wortfolge kann man suchen, indem man diese in doppelte Hochkommata einschließt. Für die Suche nach Wortbildungsmustern eignet sich das Platzhalterzeichen „*“, das eine beliebige Zahl von Zeichen beliebiger Art vertritt. Mit dem Negationsoperator „!“ kann man nach Sätzen suchen, in denen ein bestimmtes Element nicht vorkommt, z. B. passt der Suchausdruck „rümpfen && !Nase“ genau auf Sätze, in denen das Wort *rümpfen* nicht gemeinsam mit dem Wort *Nase* vorkommt.



1	1900 ZE	... Richtung einen recht einseitigen Zug auf: - sie -	findet	ihr besonderes Gepräge durch eine Anzahl von...
2	1900 ZE	...tgewerbe für den englischen Stil zu schulen. Er	fand	einen tüchtigen Helfer und geschickten Agita...
3	1900 ZE	...erte, dass eine pflichtvergessene Publicistik sich	gefunden	habe, die aus alledem dem Verwaltungsrath einen&...
4	1900 GE	... worden. Diese, hm, allgemein-kulturellen Dinge	finden	einen weiten Resonanzboden im Herzen der Nat...
5	1900 ZE	...cht, Gnade vor den Augen des Herrn Max Mauthner zu	finden	, und wurde, als er sich um das Amt eines Co...
6	1900 ZE	...en müsse. Das Kaiserjubiläums-Stadttheater aber	findet	laut einer im' Fremdenblatt 'erschienenen St...
7	1900 ZE	Erst darauf hin sah ich mir das Stück an. Ich	fand	eine sehr gut geschaute Charakterstudie aus ...
8	1900 ZE	...er Luegers Regime sehr wohl fühlen. Schließlich	findet	der Unparteiische heraus, dass es nur eine -...
9	1900 ZE	...u -, wo sie rasch und billig Rathschläge und Hilfe	finden	, wiewohl diese Anstalt nur eine fachliche M...
10	1900 ZE	...ath Professor - v. Perger -. Gestern vormittags	fand	an der technischen Hochschule eine große Ova...
11	1900 ZE	...tät gemein hat, durch gediegene Kräfte jene Pflege	finden	würden, die einer Hochschule angemessen ist....
12	1900 ZE	...n ist mit Katastrophen überreich gesegnet, und wir	finden	den Gedanken an eine Bedeckung der österreic...
13	1900 ZE	... die historischen Daten. In meinen Notizbüchern	finde	ich eine vielfach lückenhafte Sammlung der H...
14	1900 ZE	...t über schmälere -.... Und bei allen Wäschestücken	findet	sich links oben die Märke: ein großes S und ...
15	1900 ZE	...II. In Nestroys » Zwei ewige Juden und Keiner «	finde	ich (I. Act, 22. Scene) eine Stelle, die seh...
16	1900 ZE	... Versammlungsgesetzes fertig werden « wird. Ich	fand	schon damals das Ertrotzen eines Martyriums ...
17	1900 ZE	... Stück z. B. im Jubiläumstheater aufgeführt, so	findet	er vollends nicht eine Zeile am andern Tag i...
18	1900 WI	...en derselben Rasse von der Seite durchleuchtet, so	findet	man, dass die Verfeinerung wesentlich in einer...
19	1900 WI	...mhaare oft erst gegen das 10. Jahr, und bei dreien	fand	sich auf der Wirbelsäule ein Haarwirbel eben...
20	1900 WI	...chrieben, Redner hat sie aber dort ziemlich häufig	gefunden	; und zwar zu einer Zeit, wo bei den meisten...
21	1900 ZE	...Wunsiedel überwiegt der mittlere Bauernstand; dazu	findet	sich eine rührige Kleinindustrie und vereinz...
22	1900 ZE	... später nicht sehr viel höher stieg. Cumberland	fand	bald Nachahmer, die es unternahmen, die Handlungen...

Abb. 2: Ausschnitt einer Trefferliste zur Suchanfrage „finden #10 ein“

Eine nützliche Funktion in Korpusrecherchesystemen ist die Spezifikation eines Abstandsfensters. In einer DWDS-Suchanfrage kann man mit dem Abstandsoperator "#n" nach dem gemeinsamen Vorkommen von zwei Elementen suchen, die in einem Abstandsfenster von maximal n Wörtern aufeinander folgen. Zur Anfrage „finden #10 ein“ passen also alle Sätze, in denen *ein* im Abstand von maximal zehn Wörtern dem Wort *finden* folgt; einen Ausschnitt der dazu erzeugten Trefferliste findet man in Abb. 2. An der Liste erkennt man einen zentralen Unterschied zur Suche mit Google: Zum Suchwort *finden* werden alle Flexionsformen ausgegeben, also auch *findet*, *fand* und *gefunden*. Das Korpus ist also lemmatisiert und das Korpusrecherchesystem sucht automatisch nach allen Flexionsformen der eingegebenen Wortform. Diese lemmabasierte Suche ist für das Deutsche sehr vorteilhaft, schließlich müsste man andernfalls alle Flexionsformen von *finden* in eine Oder-Verknüpfung integrieren. Wie die Trefferliste in Abb. 2 zeigt, führt die standardmäßige Lemmatisierung allerdings dazu, dass auch zum Suchwort *ein* alle flektierten Formen des homographen unbestimmten Artikelworts *ein* ausgegeben werden.

Dies kann man verhindern, indem man die Suche mit dem Symbol „@“ auf eine spezielle Form einschränkt: Die Abfrage „finden #10 @ein“ sucht nach Kombinationen von allen Flexionsformen von *finden* mit exakt der Wortform *ein* (im Abstandsfenster von zehn Wörtern). Von den in Abb. 2 gezeigten Treffern würden lediglich Treffer 14

und 19 diesem Kriterium entsprechen. Wer allerdings mit dieser Abfrage nach Belegen gesucht hat, in denen das Verb *einfinden* (in getrennter Stellung) vorkommt, wird enttäuscht sein, denn beide Trefferlisten enthalten überwiegend Pseudotreffer; auch der in Abb. 2 gezeigte Ausschnitt enthält keinen einzigen Beleg für *einfinden*. Um die Präzision der Anfrage wirklich zu verbessern, muss man die Wortartenannotation nutzen; für eine solche Suche stellt das DWDS-System spezielle Operatoren bereit: Zur Abfrage „finden #10 ein with $p=PTKVZ“ passen nur Sätze, in denen die Wortform *ein* als PTKVZ (trennbarer Verbzusatz) annotiert ist. Wenn man den in Abb. 3 gezeigten Ausschnitt der Trefferliste zu dieser Anfrage mit dem Ausschnitt in Abb. 2 vergleicht, wird der positive Effekt sehr deutlich: Die Liste in Abb. 3 enthält fast nur Belege für *einfinden*; es gibt lediglich zwei Pseudotreffer: Beim (Pseudo-)Treffer 3 ist die Wortform *ein* zwar korrekt als Verbzusatz annotiert; der Verbzusatz gehört aber zum Verb *einschlagen* und nicht zum Verb *einfinden*. Beim (Pseudo-)Treffer 2 hingegen ist das Wortvorkommen kein Verbzusatz; hier liegt vermutlich ein Fehler bei der Annotation vor. Der Vergleich der beiden Trefferlisten zeigt dennoch, dass die Wortartenannotation die Präzision der Suchanfragen stark verbessert, auch wenn die automatische Zuordnung nicht in allen Fällen fehlerfrei ist.

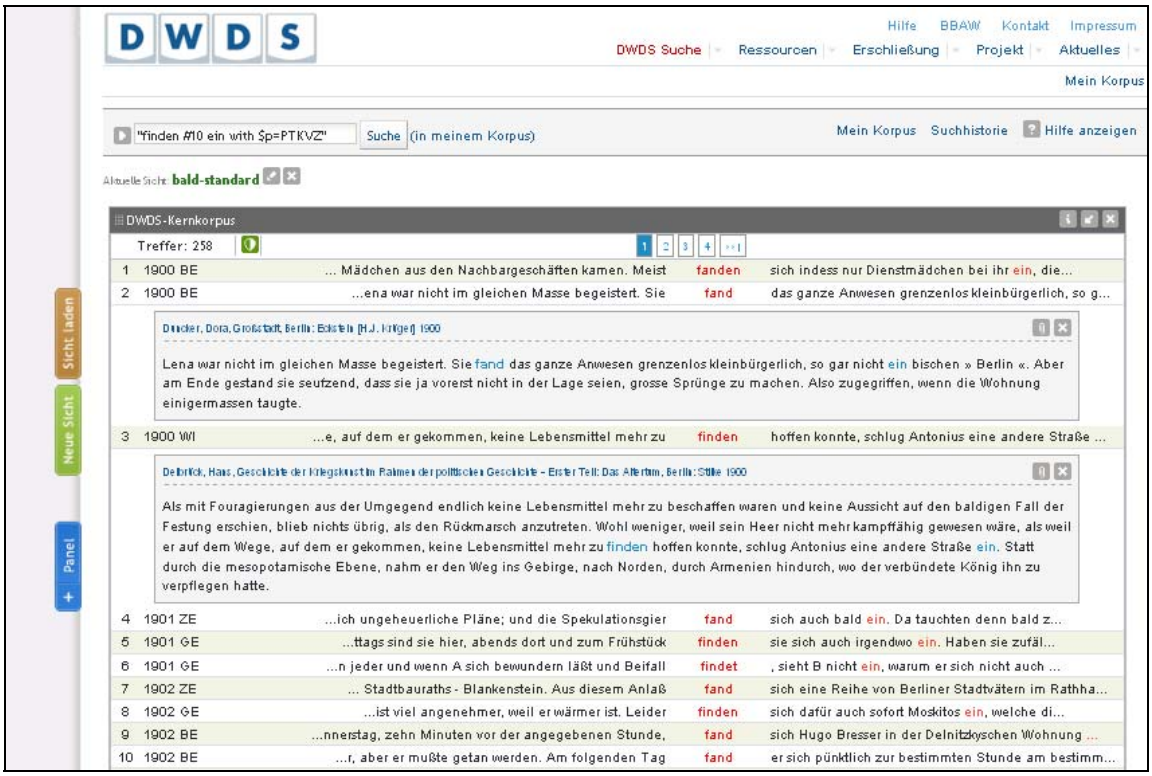


Abb. 3: Ausschnitt einer Trefferliste zur Suchanfrage „finden #10 ein with $p=PTKVZ“

Die Trefferlisten in Abb. 2 und 3 sind nach dem Prinzip „Keyword in Kontext“ (abgekürzt als KWIC) angezeigt, das in Korpusrecherchesystemen weit verbreitet ist. Die KWIC-Sicht eignet sich für die schnelle Durchsicht vieler Belege. Im DWDS-System kann man den Satzkontext und die Metadaten eines Treffers bei Bedarf per Mausklick auf eine KWIC-Zeile dazuschalten (vgl. die expandierte Anzeige der Treffer 2 und 3 in Abb. 3). Die Standardanzeige des Systems sortiert chronologisch absteigend; der Nutzer kann aber andere Optionen der Sortierung einstellen, z. B. eine Ordnung nach Satz- und Dokumentenlänge oder eine Sortierung nach dem Zufallsprinzip. Für Analysen zur Wortschatzentwicklung empfiehlt sich eine chronologische Sortierung der Treffer nach Erscheinungsdatum (vgl. chronologisch aufsteigend in Abb. 2 und 3). Die zufällige Anordnung ist nützlich, wenn für eine Suchanfrage so viele Treffer ausgegeben werden, dass man nur Teilmengen davon intellektuell analysieren kann. Da bei der zufälligen Sortierung Treffer aus unterschiedlichen Zeitabschnitten und Textsortenbereichen gemischt werden, kann man aus einer solchen Liste unkompliziert eine bearbeitbare Teilmenge an Treffern gewinnen. Für spezielle Fragestellungen bietet das Korpusrecherchesystem auch die Möglichkeit, den Suchraum auf einen Textsortenbereich (z. B. nur Belletristik) oder einen bestimmten Zeitabschnitt zu beschränken. Basis für diese Filteroptionen sind die Metadaten, die den Dokumenten des DWDS-Kernkorpus beigefügt sind. Das DWDS-Korpusrecherchesystem stellt zudem einige Statistikfunktionen bereit, die Daten, Metadaten und Annotationen des Korpus auswerten; einige davon werden in den Beispielen in Abschnitt 4 vorgestellt.

2.3 Typen von Korpora

In der korpusgestützten Sprachanalyse wird vornehmlich mit linguistisch aufbereiteten Korpora gearbeitet; diese stehen auch in diesem Artikel im Vordergrund. Unter einem linguistisch aufbereiteten Korpus verstehe ich ein digitales Korpus, das über linguistische Annotationen und Metadaten und ein Korpusrecherchesystem für die korpusgestützte Sprachanalyse verfügt. Die Verfügbarkeit linguistischer Annotationen und einer darauf zugeschnittenen Recherchesoftware unterscheidet linguistisch aufbereitete Korpora von digitalen Textarchiven wie z. B. dem Projekt Gutenberg oder digitalen Zeitungsarchiven. Diese bieten zwar oft Metadaten zu Autor, Thema und Erscheinungsdatum; die Recherchewerkzeuge sind aber eher an Themen und Inhalten orientiert, während die linguistische Suche wegen der fehlenden Lemmatisierung und Wortartenannotation vergleichsweise umständlich ist.

Die Verfügbarkeit linguistischer Annotationen und die Nutzung spezialisierter Werkzeuge unterscheidet das Arbeiten mit linguistisch aufbereiteten Korpora auch von der Nutzung von Suchmaschinen wie Google, die für die Informationsrecherche im World Wide Web konzipiert sind. Im Prinzip kann man das World Wide Web bzw. ausgewählte Teilbereiche davon als Korpus im Sinne der o. g. Definition verstehen, auch wenn die

aus dem WWW stammenden Belege natürlich mit Bedacht interpretiert werden müssen (vgl. Lemnitzer/Zinsmeister 2006:43 und Bergh/Zanchetta 2008). Für lexikographische Anwendungen sind insbesondere die Frequenzangaben interessant: Bickel (2006) zeigt beispielsweise, wie webbasierte Frequenzvergleiche im WWW die Kompetenz der Lexikographen beim Aufbau eines Varietätenwörterbuchs unterstützen können. Auch bei der computerunterstützten Untersuchung zur Wortschatzentwicklung sind Frequenzangaben aus dem WWW interessant. Die Neuwortsammlung „Die Wortwarte“, die täglich die in Tageszeitungen verwendeten Wortvorkommen gegen eine Liste bereits bekannter Wörter abgleicht und auf diese Weise semi-automatisch neue Bildungen aufspürt (vgl. [Wortwarte] und Lemnitzer 2007), gibt zu diesen neben den Belegen auch die Frequenzen aus Google an. Die Beobachtung der Frequenzentwicklung gibt Hinweise darauf, ob es sich bei der Neubildung um einen auf einen spezifischen Kontext zugeschnittenen Okkasionalismus handelt oder ob die Neubildung häufig genug verwendet wird, um als neue Wortschatzeinheit in ein künftiges Wörterbuch aufgenommen zu werden.

Von der Frequenzauswertung abgesehen, ist das WWW in seiner linguistisch nicht weiter aufbereiteten Form für die korpusgestützte Sprachanalyse nur bedingt geeignet. Deshalb arbeiten korpuslinguistische Projekte an Werkzeugen, um aus dem Web zusammengestellte Korpora linguistisch aufzubereiten. Das Projekt „ Web-as-Corpus kool ynitiative WaCKy“ entwickelt Werkzeuge und Ressourcen zur (computer)linguistischen Analyse von Daten aus dem WWW [WaCKy-Home]; für deren Nutzung ist allerdings eine computerlinguistische oder informatische Vorbildung hilfreich. Linguistisch aufbereitete Daten aus dem WWW bezeichnet man als ‚Webkorpora‘. Diese Bezeichnung ist nicht zu verwechseln mit der Bezeichnung ‚Online-Korpus‘ bzw. ‚online verfügbares Korpus’, die lediglich impliziert, dass das betreffende Korpus über eine Nutzerschnittstelle im WWW direkt zugänglich ist. Die in Abschnitt 3 beschriebenen Online-Korpora zum Deutschen sind keine Webkorpora, sondern Text- bzw. Gesprächsammlungen, in denen direkt mit einem Webbrowser recherchiert werden kann.

Die Unterscheidung zwischen Textkorpora und Gesprächskorpora orientiert sich an der medialen Realisierung der im Korpus gesammelten Sprachdaten: Textkorpora enthalten schriftlich produzierte Texte, Gesprächskorpora enthalten mündliche Gespräche, die meist in transkribierter Form vorliegen. In alignierten Gesprächskorpora sind die Transkripte mit den zugehörigen Audio- bzw. Videofiles der Gespräche verknüpft. Für die Erforschung der Kommunikation in den sozialen Netzwerken des Internets benötigt man zunehmend auch multimediale Webkorpora, die aus verlinkten Text-, Bild-, Audio- und Videodaten bestehen und sich deshalb nicht den beiden grundlegenden Kategorien ‚Text’ vs. ‚Gespräch’ zuordnen lassen (Beißwenger/Storrer 2008; Mehler 2008).

Die Unterscheidung zwischen Referenzkorpora und Spezialkorpora orientiert sich am sprachlichen Gegenstandsbereich, der durch die Korpusdokumente abgedeckt werden soll. Referenzkorpora möchten die Allgemeinsprache eines bestimmten Zeitabschnitts

repräsentieren; Spezialkorpora beschränken sich bewusst auf ausgewählte Textsortenbereiche, Autoren oder Varietäten.

Ein Leitprojekt für Referenzkorpora zu einer Nationalsprache ist das „British National Corpus“ BNC, das ca. 100 Millionen Textvorkommen mit Texten aus verschiedenen Textsortenbereichen zum britischen Englisch ab 1960 erfasst [BNC]. Die im BNC angelegten Leitlinien zur Korpuszusammenstellung und -aufbereitung waren Vorbild für ähnliche Projekte zu anderen Nationalsprachen (Amerikanisches Englisch, Russisch, Chinesisch, Tschechisch, Polnisch etc.; Beschreibungen und URLs finden sich in Xiao 2008).

Am Design des BNC orientiert sich auch das deutsche DWDS-Kernkorpus, das im Rahmen des Projekts „Digitales Wörterbuch der deutschen Sprache“ an der Berlin-Brandenburgischen Akademie der Wissenschaften aufgebaut wurde (vgl. Klein 2004; Geyken 2005). Das Kernkorpus enthält ebenfalls ca. 100 Mio. Wortvorkommen aus ca. 80.000 Dokumenten, die jeweils vier Textsortenbereichen zugeordnet sind: Gebrauchsliteratur (GE), Belletristik (BE), Wissenschaft (WI), und Zeitungen (ZE). Anders als das BNC deckt das DWDS-Kernkorpus das komplette 20. Jahrhundert ab; es eignet sich deshalb auch sehr gut für die Analyse von Wortschatzentwicklungen im 20. Jahrhundert. Das Korpus strebt an, jede Dekade des 20. Jahrhunderts mit möglichst gleich vielen Wortvorkommen abzudecken. Weiterhin sollen in jeder Dekade möglichst alle Textsortenbereiche in ausgewogenem Verhältnis vertreten sein (vgl. Geyken 2007). Nicht zuletzt wegen urheberrechtlicher Probleme konnte diese Idealverteilung bislang nur annäherungsweise umgesetzt werden; die aktuelle und die geplante Verteilung sind in den Metatexten der Online-Schnittstelle [DWDS] einsehbar.

Die nach dem Vorbild des BNC zusammengestellten Korpora streben an, mehrere Textsortenbereiche in einem möglichst ausgewogenen Verhältnis zusammenzustellen. Dieses Leitbild des ausgewogenen Korpus ist die bescheidenere Alternative zum Anspruch des repräsentativen Korpus, der in der Korpuslinguistik schon früh kritisch diskutiert wurde (zu dieser Diskussion: Lemnitzer/Zinsmeister 2006:50ff.; McEnery/Xiao/Tono 2006:13ff.). Um ein repräsentatives Korpus zu einer Nationalsprache zusammenstellen zu können, müsste man einen Gegenstand wie ‚das britische Englisch der Gegenwart‘ in seiner Gesamtheit und Zusammensetzung kennen; erst auf dieser Basis kann man eine repräsentative Stichprobe ziehen. Für eine Nationalsprache ist dies nicht realistisch. Insbesondere wäre es schwierig, die Anteile der gesprochenen Sprache zu bemessen und im richtigen Verhältnis in der Stichprobe zu berücksichtigen. Man versucht deshalb beim Design von Referenzkorpora, durch die ausgewogene Mischung verschiedener Textsortenbereiche und Zeitabschnitte dem Ideal der Repräsentativität möglichst nahe zu kommen.

Es gibt aber auch viele Korpusprojekte, in denen der Aspekt der Ausgewogenheit keine Rolle spielt, weil es vornehmlich darum geht, möglichst viele Texte eines Sprachaus-

schnitts verfügbar zu machen. Mit dem Ausdruck ‚opportunistisch zusammengestellte Korpussammlungen' werden solche Korpora von den ausgewogenen Korpora unterschieden.

Für die Zitierbarkeit von Auswertungsdaten ist es wichtig zu wissen, ob sie sich auf ein statisches Korpus beziehen, das aus einer unveränderlichen Zahl von Dokumenten besteht, oder auf ein dynamisches Korpus, das seinen Bestand im Laufe der Zeit verändert. Da auch statisch konzipierte Korpora wie das BNC oder das DWDS-Kernkorpus immer wieder neue, verbesserte Versionen generieren, empfiehlt es sich, bei einer korpusgestützten Studie nicht nur die Suchanfrage, sondern auch das Datum der Suche zu speichern, um die Ergebnisse ggf. reproduzierbar zu machen.

3 Online-Korpora zur deutschen Sprache: Überblick

An digitalen Textsammlungen im Internet oder auf CD-ROM herrscht kein Mangel, laufende kommerzielle und national geförderte Digitalisierungsprojekte werden das Angebot künftig noch vergrößern. Für die korpusgestützte Sprachanalyse sind sie wegen der fehlenden linguistischen Aufbereitung allerdings nur bedingt geeignet – wie im vorigen Abschnitt gezeigt, unterstützen linguistisch aufbereitete Korpora mit spezialisierten Recherchewerkzeugen die gezielte Suche nach sprachlichen Einheiten besser als die Suchtechnologien des Internets oder die auf thematische Recherche spezialisierten Suchwerkzeuge in Zeitungsarchiven. Die folgenden drei linguistisch aufbereiteten Online-Korpora zur deutschen Sprache sind kostenfrei verfügbar und ohne computerlinguistische Vorbildung nutzbar:

(1) Das Institut für deutsche Sprache IDS in Mannheim besitzt die größte Sammlung von Korpora geschriebener deutscher Gegenwartssprache; in ihnen kann man mit dem Korpusrecherchesystem COSMAS recherchieren [IDS-Korpora-geschrieben]. Das Korpus umfasst viele meist opportunistisch zusammengestellte Teilkorpora, die teilweise auch lemmatisiert und wortartenannotiert vorliegen. Die Nutzer können aus dem sehr großen Gesamtbestand eine zur Untersuchungsfrage passende Auswahl treffen. COSMAS verfügt über flexible Such- und Auswertungsmöglichkeiten, außerdem werden verschiedene Werkzeuge zur quantitativen Auswertung (Frequenz, Kookkurrenzprofile) online angeboten, die auch in den lexikographischen und grammatikographischen Projekten des Instituts genutzt werden. Mit Bubenhofer (o. J.) liegt eine Online-Einführung in die Korpuslinguistik vor, in der Funktionen und Anwendungsoptionen für COSMAS an praktischen Beispielen erläutert werden.

(2) Das Institut für deutsche Sprache verfügt mit dem „deutschen Spracharchiv" auch über umfangreiche Bestände transkribierter gesprochener Sprachdaten. Teile dieser Transkriptsammlung sind über die Online-Schnittstelle der „Datenbank gesprochenes Deutsch" verfügbar [IDS-Korpora-gesprochen]; einige davon sind

aligniert, d. h. die Transkripte sind mit den zugehörigen Ausschnitten der Audiodateien verlinkt.

(3) Die Berlin-Brandenburgische Akademie der Wissenschaften bietet einen kostenfreien Online-Zugang zu mehreren linguistisch aufbereiteten Korpora, die im Rahmen des Projekts „Digitales Wörterbuch der deutschen Sprache (DWDS)“ aufgebaut werden. Neben dem bereits in Abschnitt 2.3 beschriebenen DWDS-Kernkorpus stehen mehrere Zeitungskorpora (u. a. die ZEIT) sowie verschiedene Spezialkorpora zur Wahl. Die Korpora des DWDS sind gemeinsam mit digitalen Wörterbuchdaten in das lexikalische Wortinformationssystem [DWDS] integriert, das in Abschnitt 4.1 ausführlicher beschrieben wird.

Für die Recherche nach syntaktischen Fragestellungen sind syntaktisch komplett annotierte Korpora, sog. Baumbanken, eine große Hilfe. Im Prinzip erfordert der Umgang mit Baumbanken keine computerlinguistische Ausbildung; allerdings muss man für den Einstieg in die Recherche mehr Zeit einplanen als bei anderen Online-Korpora, denn man muss sich sowohl in das Suchwerkzeug als auch in das Kategoriensystem einarbeiten, das im jeweiligen Korpus für die syntaktische Annotation genutzt wird. Die am Institut für maschinelle Sprachverarbeitung (IMS) der Universität Stuttgart entwickelte Recherchesoftware „TiGerSearch“ ist ein intuitiv bedienbares, gut dokumentiertes und ansprechend gestaltetes Werkzeug, das für wissenschaftliche Zwecke kostenfrei auf verschiedenen Plattformen installiert werden kann [TiGerSearch]. Für Einsteiger bietet TiGerSearch eine graphische Abfragesprache, in der sich mit einfachen Abfragen an einem Beispielkorpus das Prinzip der Suche in Baumbanken erlernen lässt. Fortgeschrittenen Nutzern bietet die Syntax der symbolischen Abfragesprache flexible Suchoptionen.

Als Datenbasis stehen für die deutsche Gegenwartssprache mehrere Baumbanken zur Verfügung: Die an der Universität des Saarlandes aufgebaute „NEGR@“-Baumbank wurde semi-automatisch erstellt und intellektuell annotiert. Sie verfügt in ihrer aktuellen, zweiten Version über ca. 20.000 annotierte Sätze aus deutschen Zeitungstexten (Frankfurter Rundschau) [Negra-BB]. Die an der Universität Tübingen entwickelte „Baumbank des Deutschen/Schriftsprache“ ist ein syntaktisch annotiertes Korpus mit Zeitungstexten (taz) im Umfang von derzeit ca. 45.000 Sätzen [TüBa-D/Z]. Die ebenfalls in Tübingen entwickelte „Baumbank des Deutschen/Spontansprache“ ist ein Korpus manuell transliterierter spontansprachlicher Dialoge und umfasst ca. 38.000 Sätze [TüBa-D/S]. Beide Tübinger Korpora berücksichtigen neben der Konstituentenstruktur und den grammatischen Funktionen auch topologische Felder. Die am IMS der Universität Stuttgart erstellte „TiGer“-Treebank (Version 2.1) umfasst ca. 50.000 Sätze Zeitungstext (Frankfurter Rundschau) und eignet sich wegen der engen Verbindung zum TiGer-Search-Werkzeug (ein Pröbchen dieses Korpus ist dem Werkzeug beigefügt) besonders gut dazu, den Umgang mit dem Werkzeug einzuüben und sich das Potenzial

der Recherche in Baumbanken für die Sprachforschung zu erschließen [TiGer-BB]. Es ist aber gerade eine Stärke von TiGer-Search, dass auch die Formate von NEGR@ und TüBa-D/Z, sowie andere Baumbank-Standards (z.B. das Format der englischen PENN-Treebank) unterstützt werden.

Neben den genannten Ressourcen zur deutschen Gegenwartssprache gibt es noch andere Korpora, u. a. zu Varietäten und Sprachstadien des Deutschen in verschiedenen Stadien der linguistischen Aufbereitung. In Lemnitzer/Zinsmeister (2005:Kap.5) findet man einen systematischen, ausführlichen Überblick zu deutschsprachigen Korpora; Xiao (2008) beschreibt einflussreiche Korpora unterschiedlichen Typs (viele Sprachen, den Schwerpunkt bildet das Englische). Da sich die Korpuslinguistik sehr rasch entwickelt, empfiehlt es sich, bei der Suche nach spezielleren Korpora auch Online-Angebote zu konsultieren, z. B. auf dem Essener Linguistik-Server die LINSE-Rubrik zur Korpuslinguistik [Linse-Korpora] oder die Sammlung der Evaluations and Language Resources Distribution Agency (ELDA), die Korpora und lexikalische Ressourcen mit Schwerpunkt auf europäischen Sprachen distribuiert [ELDA-Korpora]. Das Institut für maschinelle Sprachverarbeitung IMS in Stuttgart pflegt eine sehr nützliche Linkliste speziell zu Baumbanken und Baumbankprojekten in vielen Sprachen [IMS-Baumbanken].

4 Digitale Korpora in Lexikographie und Phraseologie

Das Arbeiten mit Korpora hat gerade in der Lexikographie eine lange Tradition. Selbst bei gegenwartssprachlichen Wörterbüchern würde sich kein Lexikographenteam anmaßen, eine vollständige lexikographische Beschreibung allein auf der Basis der eigenen Sprachkompetenz auszuarbeiten. Vielmehr exzerpieren und analysieren seriöse Wörterbuchprojekte Belege aus Quellentexten und konsultieren andere Wörterbücher als Sekundärquellen. Unumgänglich sind Korpora für die Beschreibung älterer Sprachstufen, für die zeitgenössische Lexikographen ja keine muttersprachliche Kompetenz mitbringen. Die Wörterbuchforschung hat die Prozesse der Erarbeitung von gedruckten Wörterbüchern sehr detailliert erfasst und beschrieben (vgl. Wiegand 1998). Die folgende stark vereinfachte Skizze der „prädigitalen“ Korpusnutzung soll dazu dienen, den qualitativen Sprung deutlich zu machen, der durch die Verfügbarkeit digitaler Korpora entsteht. Beim prädigitalen Vorgehen werden aus Quellenkorpora, die eine möglichst vielfältige und ausgewogene Auswahl von Texten zum jeweils relevanten Sprachausschnitt enthalten, Belegstellen exzerpiert und in einem Belegarchiv alphabetisch nach Stichwörtern geordnet. Diese Belegarchive sind in verschiedenen Phasen des lexikographischen Prozesses wichtig: bei der Entscheidung, welche Stichwörter ins Wörterbuch aufgenommen werden, bei der Bestimmung, wie viele semantische Lesarten für ein Stichwort angesetzt werden und bei der Formulierung der lexikographischen Angaben zu Form und Bedeutung. Manche Wörterbücher integrieren auch ausgewählte Belege in

die Wörterbuchartikel; ein Beispiel hierfür findet sich im Wörterbuchartikel zu *Ampel* in Abb. 4 links oben.

Die Vorteile digitaler Korpustechnologie für lexikographische Arbeitsprozesse liegen auf der Hand: (1) Aus digitalen Korpora kann man flexibel Trefferlisten generieren; die zeit- und kostenaufwändige Exzeption von Belegen und die „Verzettelung" in prädigitale Belegzettelarchive entfällt. (2) Digital verwaltete Belege können quantitativ ausgewertet werden; insbesondere lassen sich Daten zur Frequenz und zum gemeinsamen Auftreten von Wortvorkommen (Kollokationen/Kookkurrenz) berechnen (vgl. Geyken 2004). Natürlich unterscheiden sich die Trefferlisten, die von einem Korpusrecherchesystem erzeugt werden, vom prädigitalen Zettelarchiv: Wie im zweiten Abschnitt erläutert, operiert die automatische Suche in digitalen Korpora vornehmlich über Wortformen und formbasierten Suchmustern und nicht über Lexemen in einer bestimmten Bedeutung. Linguistische Annotationen können zwar die Präzision der Suchanfragen deutlich verbessern, dennoch enthalten die automatisch erzeugten Trefferlisten oft auch Pseudotreffer, die manuell aussortiert werden müssen (vgl. die Beispiele in 2.2). Diesen Beschränkungen zum Trotz bietet bereits die aktuelle Korpustechnologie einem methodisch und technisch kompetenten Lexikographen Optionen zur Recherche und Analyse, die in einem prädigitalen Zettelarchiv nicht oder nur mit sehr hohem Zeitaufwand möglich wären.

Die Nutzung digitaler Korpora in der Lexikographie ist in der korpuslinguistischen Literatur bereits gut beschrieben. Ein englischer „Klassiker" ist Sinclair (1991), einen aktuellen Überblick geben u. a. Lemnitzer/Zinsmeister (2006:143ff.), McEnery/Xiao/Tono (2006:80ff.) und Heid (2008). Digitale Korpusrecherchesysteme, in denen man sehr flexibel nach Wortkombinationen suchen kann, sind insbesondere für die Phraseologieforschung attraktiv. Interessante Ergebnisse aus korpusgestützten Projekten zu Idiomen und Kollokationen sind u. a. dokumentiert in Moon (1998) (für das Englische) und Fellbaum (2007) (für das Deutsche). Es ist das Anliegen der folgenden Abschnitte, die Vorteile der Korpusnutzung in Lexikographie und Phraseologie an einfachen Fallbeispielen zu illustrieren.

4.1 Digitale Wörterbücher und Korpora

Digitale Medien und das Internet verändern nicht nur die Prozesse der Wörterbuchherstellung, sondern auch die dabei entstehenden lexikographischen Produkte, die als Wörterbuchportale bzw. lexikalische Informationssysteme direkt im Internet angeboten werden (vgl. Engelberg/Lemnitzer 2009; Storrer 2010). Beim Aufbau digitaler Wörterbücher müssen sich die Wörterbuchmacher nicht mehr darum bemühen, möglichst viele Informationen auf einer Druckseite unterzubringen; die lexikographischen Angaben können deshalb übersichtlicher präsentiert und durch mehr Belegbeispielangaben angereichert werden (vgl. Storrer 2001). Von dieser Option, Wörterbuchartikel um Korpus-

belege anzureichern, machen zwei digitale Wörterbuchportale zur deutschen Gegenwartssprache Gebrauch: Das eLexiko-Wörterbuch, das im Wörterbuchportal „OWID“ des Instituts für deutsche Sprache abrufbar ist [eLexiko-OWID], und das „Projekt deutscher Wortschatz“ der Universität Leipzig, das Wörterbuchartikel semi-automatisch aus digitalen Korpora und Wörterbüchern generiert [Deutscher-Wortschatz]. Beide Wörterbücher bieten außerdem Angaben zur Frequenz der Stichwörter und zu typischen Wortverbindungen (Kollokationen, Kookkurrenzen), die automatisch aus den zugrunde liegenden Korpusdaten generiert werden: Das eLexiko-Wörterbuch ordnet alle Stichwörter einer Frequenzschicht zu und verlinkt diese mit automatisch erzeugten Kookkurrenzprofilen. Das Projekt deutscher Wortschatz gibt zu jedem Stichwort an, wie häufig dieses im zugrunde liegenden Korpus belegt ist, und ordnet es einer Häufigkeitsklasse zu, die relativ zur Häufigkeit der hochfrequenten Wortform „der“ berechnet wird. Außerdem werden typische Kollokationspartner aufgelistet und als Netzgraph dargestellt.

Einen Schritt weiter gehen digitale lexikalische Informationssysteme: Sie integrieren Wörterbuch- und Korpusressourcen durch eine Nutzeroberfläche, mit der man sowohl in Wörterbüchern als auch in Korpora recherchieren kann. Für das Deutsche wird ein solches System für den DWDS entwickelt (Klein 2004; Geyken 2005); auf die Funktionalität dieses Systems beziehen sich auch die folgenden Fallbeispiele.

Zentral für den Umgang mit der DWDS-Nutzeroberfläche sind das Konzept der Sichten und das Konzept der Panels: Als 'Sicht' bezeichnet man eine Kombination von Ressourcen (Wörterbücher, Korpora, Statistikwerkzeuge), mit der ein Nutzer arbeiten kann. Jede Ressource wird in einem als 'Panel' bezeichneten Arbeitsfenster angezeigt, das bei Bedarf vergrößert werden kann. Wer auf der Startseite ein Suchwort, z. B. das Wort *Ampel*, eingibt, erhält die in Abb. 4 gezeigte Standardsicht mit der folgenden Panel-Kombination:

(1) Das DWDS-Wörterbuch (Panel oben links) basiert inhaltlich auf dem „Wörterbuch der deutschen Gegenwartssprache“ [WDG], einem 6-bändigen Printwörterbuch, das von 1952 bis 1977 auf der Basis eines umfangreichen Quellkorpora erarbeitet wurde (vgl. Malige-Klappenbach 1986). Im Rahmen des DWDS-Projekts wurde dieses Wörterbuch digitalisiert, strukturell aufbereitet, durch vertonte Ausspracheangaben ergänzt und an die neue Rechtschreibung angepasst.

(2) Das Etymologische Wörterbuch (Panel oben rechts) ist auf Informationen zur Wortgeschichte spezialisiert. Es basiert auf der zweiten Auflage des „Etymologischen Wörterbuchs des Deutschen“ [Etym-WB], das in den 80er Jahren von einer Lexikographengruppe unter der Leitung von Wolfgang Pfeifer erstellt und im Rahmen des DWDS-Projekts digital aufbereitet wurde.

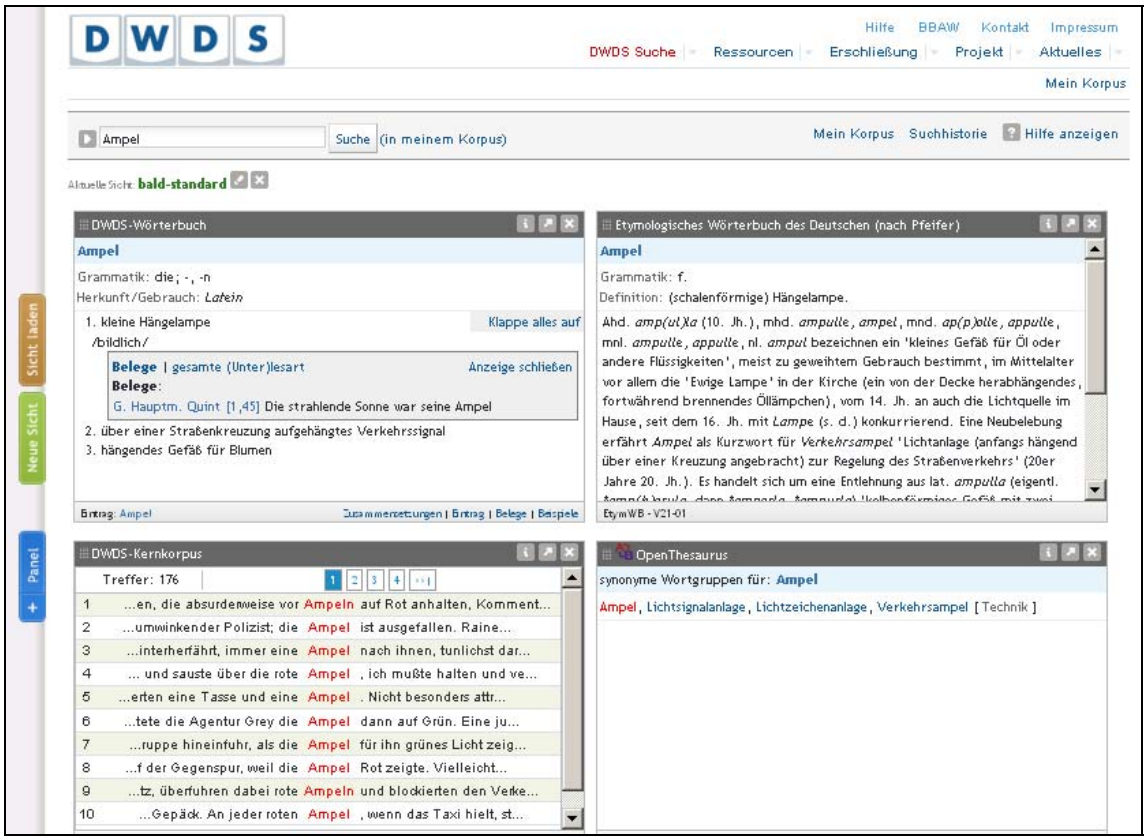


Abb. 4: Standardsicht im DWDS-System zum Suchwort *Ampel*

(3) Der OpenThesaurus (Panel unten rechts) listet Synonyme und sinnverwandte Wörter. Die Einträge des von Daniel Naber initiierten kollaborativen Wörterbuchprojekts sind im DWDS-System als externe Ressource eingebunden.

(4) Im DWDS-Kernkorpus (Panel unten links) kann man mit der in Abschnitt 2.3 beschriebenen Abfragesprache gezielt nach Suchwörtern und Suchmustern recherchieren. Wegen seiner ausgewogenen Streuung über die Dekaden des 20. Jahrhunderts und über Textsortenbereiche hinweg eignet sich dieses Korpus sehr gut dazu, Entwicklungen und Veränderungen im Wortschatz des 20. Jahrhunderts zu untersuchen.

Neben dieser Standardsicht bietet das DWDS-System weitere vordefinierte Sichten mit Kombinationen von Korpora, Korpusstatistiken und Wörterbüchern an. In den folgenden Beispielen verwenden wir zusätzlich zur Standardsicht das Zeitungskorpus der ZEIT, die Wortverlaufstatistik zum DWDS-Kernkorpus (vgl. Abb. 5) und das Statistikwerkzeug 'Wortprofil' (vgl. Abb. 6); diese und weitere Ressourcen kann man im DWDS-System in Panels dazuschalten. Registrierte Nutzer können Panelkombinationen dauerhaft als nutzerspezifische Sichten speichern. Der Aufwand für die kostenlose Registrierung lohnt sich nicht nur, weil die Definition eigener Sichten für linguistische Untersuchungsfragen oft die effizienteste Option ist, sondern weil registrierte Nutzer zudem die Möglichkeit haben,

eigene Belegsammlungen anzulegen, nach Kategorien zu klassifizieren und in einer späteren Sitzung unter dem Menüpunkt 'Mein Korpus' wieder abzurufen. Für den Einstieg in die korpusgestützte Sprachanalyse stehen damit rudimentäre Funktionen eines lexikographischen Arbeitsplatzes direkt online zur Verfügung. Die folgenden einfachen Fallbeispiele sollen illustrieren, wie die Ressourcenkombination für korpusgestützte Untersuchungen zum deutschen Wortschatz genutzt werden kann.

4.2 Frequenzinformationen und Frequenzverläufe: Analysebeispiel *Streß/Stress*

Früher hatte man weniger Stress! Ob diese oft gehörte Behauptung stimmt, kann man sicher nicht durch eine Korpusanalyse klären. Allerdings zeigt die Recherche im DWDS-Kernkorpus, dass das Suchwort *Streß* erst seit den 70er Jahren belegt ist. Am automatisch generierten Frequenzverlaufsdiagramm, das zu den 86 Treffern im DWDS-Korpus auf der Basis der Metadaten erstellt wird (vgl. Abb. 5), lässt sich weiterhin ablesen, dass das Wort zunächst überwiegend in Gebrauchstexten und wissenschaftlicher Fachliteratur vorkommt, ab den 90er Jahren aber zunehmend auch in der Belletristik und in Zeitungstexten verwendet wird. Bei der relativ geringen Treffermenge muss man diese Verteilung über die Textsortenbereiche hinweg sehr vorsichtig bewerten. Wenn man die überschaubare Trefferliste intellektuell analysiert, kann man jedoch sehr gut erkennen, wie sich das aus der Fachsprache der Psychologie stammende englische Lehnwort auch in nicht-fachsprachlichen Kontexten etabliert hat und wie sich dabei neue alltagssprachliche Formulierungsmuster und Kollokationen ausgebildet haben (z. B. *Streß haben/machen, in Streß kommen/geraten, voll/total im Streß sein, etwas artet in Streß aus* etc.).

Die Trefferliste zur Anfrage *Streß* enthält keinen Pseudotreffer; listet aber nicht alle relevanten Belege für das Lexem, denn dieses kommt auch in der Schreibvariante *Stress* vor, also in der regelkonformen Schreibvariante nach der Orthographiereform. Mit der Abfrage „Streß || Stress“ (vgl. Abschnitt 2.2) kann man nach beiden Varianten suchen, die Trefferzahl auf 106 erhöhen und eine interessante Beobachtung zur Verteilung der beiden Schreibvarianten machen: Die meisten Belege zur Schreibvariante *Stress* stammen aus den Jahren 1971–1976; zunächst wurde also die englische Schreibform auch im Deutschen verwendet. Danach überwiegt die Schreibvariante *Streß*, die bis zur Orthographiereform 1998 regelkonform war; diese Schreibung ist auch noch in Texten belegt, die in den 90er Jahren, also nach der Orthographiereform, erschienen sind. Das Beispiel zeigt generell, dass es für eine vollständige Trefferausbeute im Kernkorpus oft erforderlich ist, alle im 20. Jahrhundert zulässigen Schreibvarianten zu kombinieren; im DWDS-Wörterbuch sind die vor und nach der Reform zulässigen Varianten zu allen Stichwörtern verzeichnet.

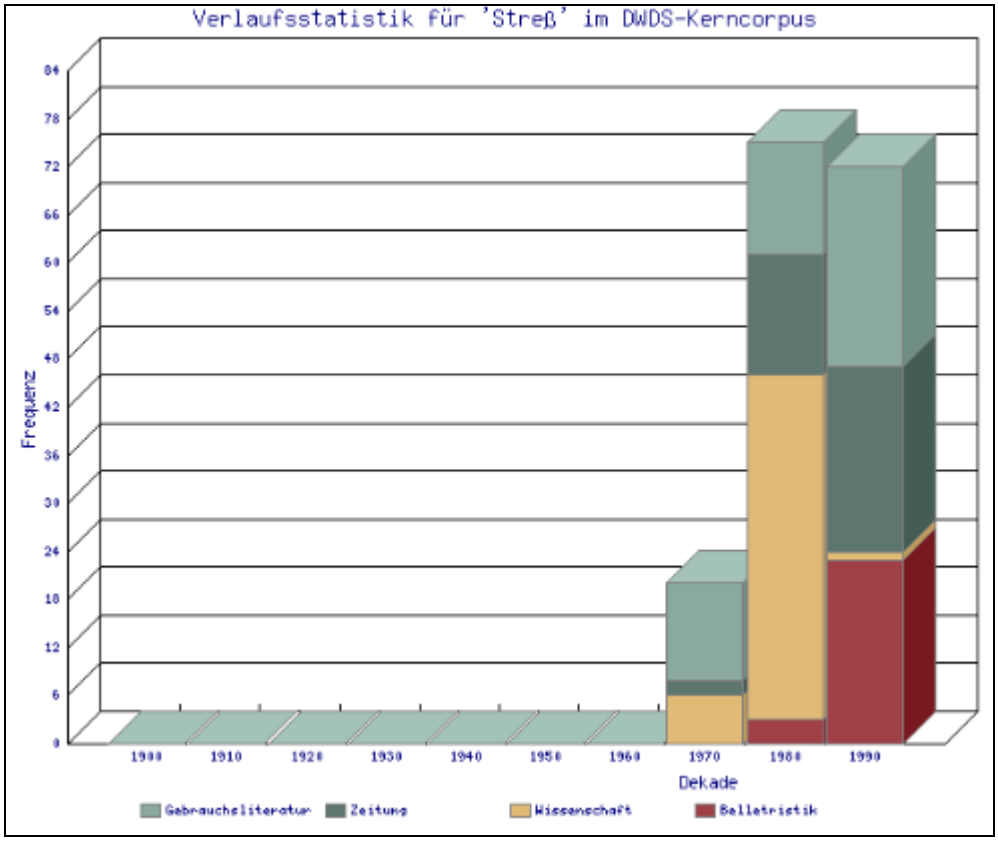


Abb. 5: Frequenzverlaufsdiagramm zum Suchwort *Streß* im DWDS-System

Am Beispielwort *Streß/Stress* kann man auch sehen, dass sich das DWDS-Korpus wegen seiner ausgewogenen Textauswahl zwar sehr dazu eignet, Sprachwandelprozesse im 20. Jahrhundert zu verfolgen, dass man die Datierung der Erstbelege aber vorsichtig interpretieren muss. Die beiden ersten Belege für die Varianten *Streß* und *Stress* im Kernkorpus stammen aus dem Jahr 1971. Das Etymologische Wörterbuch datiert die Übernahme des Lehnworts *Stress* aber bereits auf die 50er Jahre. Tatsächlich finden sich im Zeitungskorpus der ZEIT, das im DWDS-System als Panel hinzugefügt werden kann, sieben Belege aus dem Jahre 1958 und über dreißig weitere Belege aus Jahrgängen zwischen 1960 und 1970, die meisten davon in Artikeln zu medizinischen, biologischen oder psychologischen Themen. Es empfiehlt sich also, im Zweifelsfall den Datierungen im Etymologischen Wörterbuch zu vertrauen, zumindest solange man keine früheren Belege in den Korpora findet.

4.3 Bedeutungsentwicklung und Bedeutungsveränderung: Analysebeispiel *Ampel*

In Abschnitt 2.3 wurde erläutert, dass aktuell verfügbare große Korpora nicht semantisch annotiert sind, weshalb man nicht direkt nach speziellen semantischen Lesarten für

ein Lexem suchen kann. Aus diesem Grund lassen sich auch Frequenzen in Korpora nur „semantisch blind“ berechnen, was dazu führt, dass auch die automatisch generierten Frequenzverlaufsdiagramme nicht zwischen verschiedenen semantischen Lesarten einer Wortform differenzieren. Wenn man zum Suchwort *Ampel* ein Frequenzverlaufsdiagramm erzeugt, kann man zwar erkennen, dass die Wortform über das gesamte 20. Jahrhundert vor allem in der Belletristik und in der Gebrauchsliteratur belegt ist. Die formorientierte Frequenzzählung gibt aber keine Hinweise auf Verschiebungen in der Gebräuchlichkeit der drei semantischen Lesarten, die im DWDS-Wörterbuch zu diesem Stichwort verzeichnet sind: (1) 'Hängelampe', (2) 'Verkehrssignal', (3) 'Blumengefäß' (vgl. Abb. 4, Panel links oben).

Erst die intellektuelle Durchsicht der 176 Treffer zum Suchwort *Ampel* im DWDS-Kernkorpus bringt zum Vorschein, dass in der ersten Hälfte des 20. Jahrhunderts die Lesart (1) ‚Lampe‘ dominiert, während sich in der zweiten Hälfte fast nur noch Belege für die Lesart (2) ‚Verkehrssignal‘ finden. Die frühesten Belege im Kernkorpus für die Lesart ‚Verkehrssignal‘ stammen aus einem Text von Kurt Tucholsky aus dem Jahre 1933. Dieser Text enthält drei von vier Belegen für diese Lesart, die in den insgesamt hundert Treffern des Zeitabschnitts 1900–1956 zu finden sind; der vierte stammt aus dem Jahre 1951. Die restlichen 96 Treffer vor 1956 belegen ganz überwiegend die Lesart ‚Lampe‘, einige die Lesart ‚Blumengefäß‘. Nach 1956 verändert sich die Verteilung sehr rasch: In den insgesamt 76 Treffern aus dem Zeitraum 1956–1999 ist die Lesart ‚Lampe‘ nur noch zweimal belegt (1964 und 1977), die restlichen 74 Treffer belegen nur noch die Lesart ‚Verkehrssignal‘. Ein aktuelles Wörterbuch würde auf dieser Grundlage vermutlich die Lesart ‚Lampe‘ als ungebräuchlich markieren, um Missverständnisse bei der Textproduktion von Nicht-Muttersprachlern zu vermeiden.

Schwieriger ist es, Aussagen über die Bedeutung (3) (= *Ampel* als 'Blumengefäß') zu treffen, die sich vermutlich als Kurzform aus dem Kompositum *Blumenampel* entwickelt hat. Zwar kann man nachweisen, dass diese Lesart schon in den ersten Dekaden des 20. Jahrhunderts bekannt war; allerdings ist die Beleglage auch hier sehr dünn: Man findet insgesamt nur 16 Belege, acht davon stammen aus demselben Text (Paul Scheerbarts „Lesabéndio“). Nach 1956 ist die Bedeutung ‚Blumengefäß‘ im Kernkorpus nicht mehr belegt; wegen der niedrigen Frequenz in der ersten Jahrhunderthälfte sollte man daraus aber keinesfalls ableiten, dass diese Lesart nicht mehr gebräuchlich ist.

Generell sollte der Befund, dass eine Wortform oder eine Lesart im Korpus nicht belegt ist, nicht als Nachweis dafür interpretiert werden, dass die betreffende Lesart oder Wortform in der untersuchten Zeit noch nicht existiert hat. Die Neubedeutung von *Ampel* als Bezeichnung einer Koalition bundesdeutscher Parteien war schon in den 90er Jahren bekannt, auch wenn sie im DWDS-Kernkorpus nicht belegt ist. Auch in diesem Fall lohnt es sich, ergänzend im laufend aktualisierten Zeitungskorpus der ZEIT zu recherchieren: Dort findet man mehrere Belege aus den 90er Jahren, der früheste stammt aus

dem Jahr 1991. Das Kompositum *Ampelkoalition*, aus dem die Neubedeutung vermutlich durch Kurzwortbildung entstanden ist, wird in diesem Korpus erstmals 1988 verwendet. Durch die Analyse der umfangreichen Trefferliste im ZEIT-Korpus kann man verfolgen, wie sich die Neubedeutung semantisch ausdifferenziert (*schwarze Ampel*, *Schwampel*) und auch immer häufiger gebraucht wird: Im Jahrgang 2009 des ZEIT-Korpus aktualisieren bereits 79 der 143 Treffer die Neubedeutung ‚Ampelkoalition'; die übrigen Treffer 64 belegen die Lesart ‚Verkehrssignal'; kein einziger Beleg findet sich für die Lesarten ‚Lampe' oder ‚Blumengefäß'.

4.4 Typische Umgebungen/Kollokationen: Analysebeispiel *zeitigen*

Deutsche Muttersprachler haben meist ein gutes Gefühl dafür, welche Lexeme sich miteinander kombinieren lassen. Beispielweise wissen sie, dass man *einen Brand legen* und etwas *in Brand setzen* kann, dass aber die Verbindungen *einen Brand setzen* oder *in Brand legen* aber ungebräuchlich sind. Deutschlerner müssen solche kombinatorischen Präferenzen, man spricht auch von 'Kollokationen' oder 'Kookkurrenzen', die einzelsprachspezifisch sind und sich auch nicht aus der Bedeutung der kombinierten Lexeme ableiten lassen, oft im Wörterbuch nachschlagen. Korpuslinguistik und Lexikographie experimentieren seit längerem mit statistischen Verfahren, um Kollokationen bzw. Kookkurrenzen aus Korpusdaten zu gewinnen und für die lexikographische Sprachbeschreibung nutzbar zu machen (vgl. Lemnitzer/Zinsmeister 2006:145ff.; McEnery/Xiao/Tono 2006:208ff.). Auf solchen statistischen Verfahren basiert auch das Wortprofil im DWDS-System, das Kollokationen aus dem DWDS-Kernkorpus und dem ZEIT-Korpus ermittelt, nach syntaktischen Umgebungen klassifiziert und mit entsprechenden Korpusbelegen verknüpft.

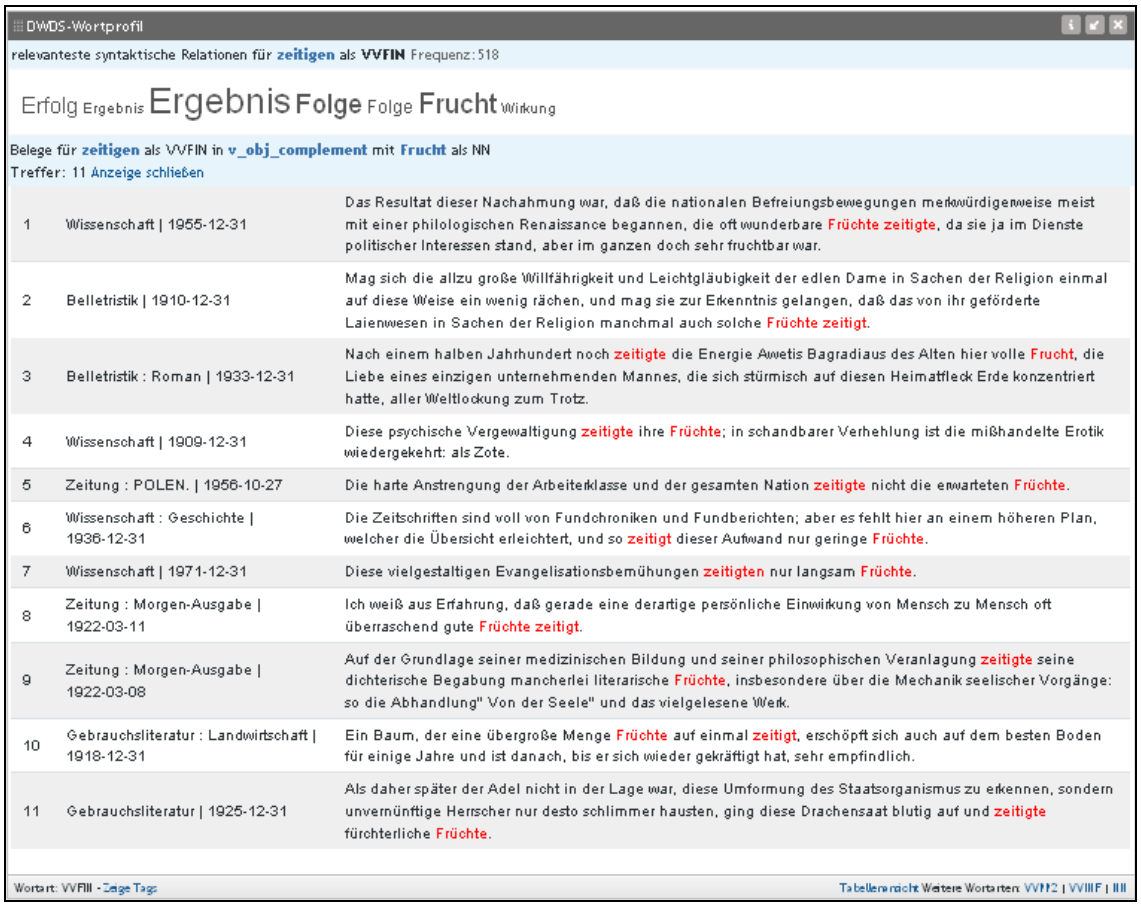


Abb. 6: Wortprofil und Belege zu *zeitigen* im DWDS-Kernkorpus

Was man aus einem solchen Profil entnehmen kann, möchte ich am Beispiel des Wortprofils für das Verb *zeitigen* illustrieren (vgl. Abb. 6). Im DWDS-Wörterbuch wird die Hauptbedeutung von *zeitigen* mit dem Synonym *hervorbringen* beschrieben. Gerade weil diese Bedeutung der gehobenen Stilschicht zugeordnet ist, kann man sich vorstellen, dass auch muttersprachliche Schreiber unsicher sind, welche Nomina bei diesem Verb als Subjekt bzw. Akkusativkomplement in Frage kommen. Bei solchen Unsicherheiten ist es möglich, sich im Wortprofil typische nominale Umgebungen anzeigen lassen, wobei die internettypische Darstellung als „Wolke“ die Kollokationspartner mit hohen Werten größer angezeigt als die mit niedrigeren Werten (in der alternativ verfügbaren Tabellensicht kann man auch die genauen Werte einsehen). Im Gegensatz zu ähnlichen Funktionen in anderen digitalen Wörterbüchern, z. B. den sehr ausführlichen Kollokationsinformationen in den Artikeln des Projekts deutscher Wortschatz [Deutscher-Wortschatz] der Universität Leipzig, sind die Wortprofile des DWDS-Systems mit den zugrunde liegenden Korpustreffern und ihren Metadaten verknüpft. Abb. 6 unten zeigt beispielsweise die Trefferliste zum Kollokationspartner *Frucht* als Akkusativkomplement. Die Durchsicht dieser Belege macht sehr schön deutlich, wie die im

DWDS-Wörterbuch aufgeführte regional markierte Lesart von *zeitigen* (österr.: 'reif werden') als lexikalisierte Metapher in der abstrakten Hauptbedeutung fortlebt.

5 Fazit und Ausblick

Linguistisch aufbereitete digitale Korpora bieten vielfältige Möglichkeiten, authentische Sprachdaten quantitativ und qualitativ zu analysieren. Die einfachen Analysebeispiele in Abschnitt 4 geben hier nur einen ersten Einblick, was man bereits ohne computerlexikographische Ausbildung in Online-Korpora entdecken kann. Die einfachen Beispiele dürften aber auch bereits deutlich gemacht haben, dass die Korpusdaten umsichtig interpretiert werden müssen, dass also digitale Korpustechnologie die lexikographische Arbeit nicht ersetzt, sondern unterstützt und ergänzt. Zentrale Fragen in lexikographischen Arbeitsprozessen wie

– Welche Lexeme werden als Stichwörter aufgenommen?
– Wie viele Lesarten setzt man für ein Stichwort an?
– Was sind typische und was sind ungewöhnliche Verwendungskontexte?

müssen auf der Basis der sachkundigen Auswertung von Korpusdaten beantwortet werden. Sachkundig bedeutet einerseits, dass man den Quellenwert der Korpusbelege richtig einzuschätzen weiß; diese Kompetenz war bereits für das prädigitale Auswerten von Belegzettelarchiven wichtig. Sachkundig heißt andererseits auch, dass man mit den Standards und Verfahren der linguistischen Aufbereitung von Annotationen vertraut ist und deren Möglichkeiten und Grenzen einschätzen kann. Wie in Abschnitt 2 erläutert, erfolgt die linguistische Aufbereitung (Lemmatisierung, Wortartenannotation etc.) in großen digitalen Korpora mit automatischen Verfahren und ist deshalb nicht fehlerfrei. Wer häufiger mit Korpora arbeitet, wird allerdings schnell Strategien entwickeln, mit denen sich die Menge der Pseudotreffer reduzieren lässt. Die Korpuslinguistik arbeitet an der Verbesserung der Verfahren und an Werkzeugen, mit denen man sehr große Treffermengen für hochfrequente Wörter lexikographisch auswerten lassen kann; ein bekanntes Beispiel ist die für das Englische entwickelte Sketch Engine [Sketch Engine]. Es wird eine spannende Aufgabe der nächsten Lexikographen-Generation sein, den Nutzwert solcher Werkzeuge in konkreten Projekten zu testen und/oder zu optimieren. Bereits jetzt bietet die aktuelle Korpustechnologie einem methodisch und technisch kompetenten Lexikographenteam Möglichkeiten der lexikologischen Recherche und Bearbeitung, wie sie in einem prädigitalen Zettelarchiv nicht oder nur mit erheblich höherem Zeitaufwand möglich wäre. Neue computerlexikographische Funktionen – z. B. die Suche nach interessanten und ungewöhnlichen Belegen, die Entdeckung und Entwicklung von Neubedeutungen, das Aufspüren von Metaphern – sollten im Dialog zwischen Lexikographie und Korpuslinguistik entwickelt und in konkreten Wörterbuchprojekten evaluiert werden. Die korpusgestützte Lexikographie wird damit sicherlich in nächster Zeit ein sehr spannendes, interdisziplinäres Betätigungsfeld für Linguis-

ten, die Spaß am Umgang mit Computern und an der empirischen Erforschung von Sprache haben.

6 Erwähnte Online-Ressourcen und Wörterbücher

6.1 Online-Ressourcen [Letzter Zugriff: 12.06.2010]

[BNC]: http://www.natcorp.ox.ac.uk
British National Corpus BNC Online

[Deutscher-Wortschatz]: http://wortschatz.uni-leipzig.de/
„Projekt deutscher Wortschatz“ (PdW), Universität Leipzig

[DWDS]: http://www.dwds.de
„Digitales Wörterbuch der deutschen Sprache“ an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW)

[ELDA-Korpora]: http://www.elda.org
Korpussammlung mit Schwerpunkt auf europäischen Sprachen der Evaluations and Language Resources Distribution Agency (ELDA)

[eLexiko-OWID]: http://www.owid.de/elexiko_/index.html
Online-Wörterbuch zur deutschen Gegenwartssprache am Institut für deutsche Sprache IDS in Mannheim

[IDS-Korpora-geschrieben]: http://www.ids-mannheim.de/kt/projekte/korpora/
Überblick über die Korpora zum geschriebenen Deutsch am Institut für deutsche Sprache IDS in Mannheim

[IDS-Korpora-gesprochen]: http://www.ids-mannheim.de/kt/projekte/korpora/archiv.html
Überblick über die Korpora zum gesprochenen Deutsch (deutsches Spracharchiv) am Institut für deutsche Sprache IDS in Mannheim

[IMS-Baumbanken]: http://www.ims.uni-stuttgart.de/projekte/TIGER/related/links.shtml
Linkliste zu Baumbanken und Baumbankprojekten am Institut für maschinelle Sprachverarbeitung IMS in Stuttgart

[Linse-Korpora]: http://www.linse.uni-essen.de/inlink/index.php?sid=793965326&t=sub_pages&cat=23
Rubrik zu Korpora und Korpuslinguistik am Essener Linguitik-Server „LINSE“ (Universität Duisburg-Essen)

[Negra-BB]: http://www.coli.uni-sb.de/sfb378/negra-corpus/negra-corpus.html
Die deutsche Baumbank „NEGR@“ (Computerlinguistik, Universität des Saarlandes)
[Sketch Engine]: http://www.sketchengine.co.uk
Homepage der Lexical Computing Ltd. (Adam Kilgarriff)

[STTS]: http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html
Online-Informationen zum Stuttgart-Tübingen TagSet zur Wortartenannotation

[TiGer-BB]: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
Die deutsche Baumbank „TiGer“ (IMS Stuttgart)

[TiGerSearch]: http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
Kostenfreies Recherchewerkzeug für Baumbanken (IMS Stuttgart)

[TüBa-D/S]: http://www.sfs.uni-tuebingen.de/tuebadz.shtml
Tübinger Baumbank des Deutschen / Spontansprache

[TüBa-D/Z]: http://www.sfs.uni-tuebingen.de/tuebads.shtml
Tübinger Baumbank des Deutschen / Schriftsprache

[WaCKy-Home]: http://wacky.sslmit.unibo.it
Homepage der „Web-as-Corpus kool ynitiative“ (WaCKy)

[Wortwarte]: http://www.wortwarte.de
„Die Wortwarte“: Laufend aktualisierte Neuwortsammlung von Lothar Lemnitzer

6.2 Wörterbücher

[WDG] Klappenbach, R. / Steinitz, W. (Hg.) (1964-1977): *Wörterbuch der deutschen Gegenwartssprache (WDG).* 6 Bände. Berlin:Akademie-Verlag.

[Etym-WB] Pfeifer, W. (1997): *Etymologisches Wörterbuch des Deutschen*. 2. Aufl. München: dtv.

7 Literaturverzeichnis

Atkins, B.T.S. / Fillmore, Ch.J. / Johnson, C.R. (2003): Lexicographic relevance: Selecting information from corpus evidence. In: *International Journal of Lexicography* 16(3): 251–280.

Beißwenger, M. / Storrer, A. (2008): Corpora of computer-mediated communication. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 292–308.

Bergh, G. / Zanchetta, E. (2008): Web linguistics. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 309–328

Bickel, H. (2006): Das Internet als linguistisches Korpus. In: *Linguistik online* 28. <www.linguistik-online.com/28_06/bickel.html> [Letzter Zugriff: 17.6.2010]

Bubenhofer, N. (o.J.): *Einführung in die Korpuslinguistik. Praktische Grundlagen und Werkzeuge*. <www.bubenhofer.com/korpuslinguistik/kurs/> [Letzter Zugriff: 17.6.2010]

Engelberg, St. / Lemnitzer, L. (2009): *Lexikographie und Wörterbuchbenutzung*. 4. Aufl. Tübingen: Stauffenburg.

Fellbaum, Ch. (Hg.) (2007): *Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies*. London: Continuum Press.

Geyken, A. (2004): Korpora als Korrektiv für einsprachige Wörterbücher. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 136: 72–100.

Geyken, A. (2005): *Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS)*. Berlin: BBAW.

Geyken, A. (2007): The DWDS corpus: A reference corpus for the german language of the 20th century. In: Fellbaum, Ch. (Hg.): *Collocations and Idioms. Corpus-based Linguistic and Lexicographic Studies*. London: Continuum Press, 23–40.

Heid, U. (2008): Corpus linguistics and lexicography. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 131–153.

Klein, W. (2004): Vom Wörterbuch zum Digitalen Lexikalischen System. In: *Zeitschrift für Literaturwissenschaft und Linguistik* 136: 10–55.

Lemnitzer, L. / Zinsmeister, H. (2006): *Korpuslinguistik: Eine Einführung*. Tübingen: Narr.

Lemnitzer, L. (2007): *Von Aldianer bis Zauselquote. Neue deutsche Wörter, woher sie kommen und wofür wir sie brauchen*. Tübingen: Narr.

Lüdeling, A. / Kytö, M. (2008) (Hg.): *Corpus Linguistics. An International Handbook*. 1. Bd. Berlin: Mouton de Gruyter.

Lüdeling, A. / Kytö, M. (2009) (Hg.): *Corpus Linguistics. An International Handbook*. 2. Bd. Berlin: Mouton de Gruyter.

Malige-Klappenbach, H. (1986): *Das Wörterbuch der deutschen Gegenwartssprache: Bericht, Dokumentation und Diskussion*. Tübingen: Niemeyer.

Moon, R. (1998): *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford: Oxford University Press.

McEnery, T. / Xiao, R. / Tono, Y. (2006): *Corpus-Based Language Studies – an advanced resource book*. London: Routledge.

Mehler, A. (2008): Large text networks as an object of corpus-linguistic studies. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 328–383.

Rayson, P. / Stevenson, M. (2008): Sense and semantic tagging. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 564–578.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Storrer, A. (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, I. / Schröder, B. / Storrer, A. (Hg.): *Chancen und Perspektiven computergestützter Lexikographie*. Tübingen: Niemeyer, 88–104.

Storrer, A. (2006): Funktionen von Nominalisierungsverbgefügen im Text. Eine korpusbasierte Fallstudie. In: Prost, K. / Winkler, E. (Hg.): *Von der Intentionalität zur Bedeutung konventionalisierter Zeichen*. Tübingen: Narr, 147–178.

Storrer, A. (im Druck): Deutsche Internet-Wörterbücher: Ein Überblick. In: *Lexicographica. International Annual for Lexicography / Revue Internationale de Lexicographie / Internationales Jahrbuch für Lexikographie* 27 (2010).

Xiao, R. (2008): Well-known and influential corpora. In: Lüdeling, A. / Kytö, M. (Hg.): *Corpus Linguistics*. 1. Bd. Berlin: Mouton de Gruyter, 383–457.

Wiegand, H.E. (1998): *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1. Bd. Berlin: de Gruyter.