Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten

Hans Bickel/Markus Gasser/Annelies Häcki Buhofer/Lorenz Hofer/Christoph Schön (Basel)

Abstract

The SWISS TEXT CORPUS (CHTK) has made it its goal to extensively document the German language of the 20th century in Switzerland. In this way, and in its parallel function as a sub-corpus of the Corpus C4, that will consist of 20 million text words (tokens) each from Germany, Austria, Italy/South Tirol and, as already said, Switzerland, it represents a classical reference corpus both for the standard German language in Switzerland as well as in the entire German-speaking area of Western Europe. A reference corpus should meet the requirement of comprehensively depicting the central repertoire of a language, i.e. the generally used vocabulary of this language, which is why questions of corpus structure and general planning (corpus design) play a decisive role (cf. Lemnitzer/Zinsmeister (2006: 106), where the type of the reference corpus is contrasted with the special corpus). Four and a half years after the start of the project, the SWISS TEXT CORPUS was made available to the general public in April 2009, as a research instrument. The following article outlines in brief the history of this research project and deals with fundamental and specific decisions that had to be made in the design of such a reference corpus, and with how the CHTK is compiled. Together with a concluding overview of some retrieval and analysis options offered by the CHTK, this article also provides an overview of the potential of this new research instrument and supplies the background knowledge required to work with the CHTK. For reasons of space, the methods of working, the corpus-driven approaches, cannot be thematised here (cf. Bubenhofer 2008, 2006).

1 Geschichte und Forschungsrahmen

Die Initiative für ein neues Korpus deutschsprachiger Texte ging von Berlin aus. Unter dem Titel *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts* (*DWDS*)[1] wurde im Jahr 2000 eine Projektgruppe gebildet, die das Ziel hatte, ein korpusbasiertes digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts zu erstellen, wobei der Fokus für die erste Phase stark auf das grundlegende Arbeitsinstrument eines solchen Wörterbuchs, eben das elektronische Korpus, ausgerichtet war.[2]

Von Anfang an war geplant, dass es sich dabei nicht um ein rein bundesdeutsches, also auf die Bundesrepublik Deutschland begrenztes Projekt handeln sollte, sondern dass die anderen Varietäten des Deutschen im Korpus ebenso vertreten sein sollten. Um diesen Anspruch auch auf organisatorischer Ebene umzusetzen, wurden Kooperationspartner aus den anderen deutschsprachigen Ländern gesucht. Im März 2001 konnte zwischen der Berlin-Brandenburgischen Akademie der Wissenschaften, der Österreichischen Akademie der Wissen-

[1] Cf. Berlin-Brandenburgische Akademie der Wissenschaften.

[2] Einen Überblick über das DWDS-Projekt gibt Klein (2004).

schaften und der Schweizerischen Akademie der Geisteswissenschaften eine Vereinbarung geschlossen werden, in welcher sich die Akademien zum Aufbau eines gemeinsamen Korpus verpflichteten.[3] Während in Wien genau auf diesen Zeitpunkt hin das *Austrian Academy Corpus*[4] als Partnerorganisation gegründet werden konnte, gestaltete sich die Suche nach einem geeigneten Partner in der Schweiz schwieriger – nicht zuletzt, weil die Finanzierung nicht über die Akademie gewährleistet werden konnte, sondern als normales Forschungsgesuch die Evaluationsprozeduren des Nationalfonds zu durchlaufen hatte.

Am 1. Oktober 2004, mit dreijähriger Verspätung auf die Partnerorganisationen, konnte das CHTK unter der Leitung von Annelies Häcki Buhofer seine Arbeit am Deutschen Seminar der Universität Basel aufnehmen. Das CHTK konnte mit durchschnittlich rund 200 Stellenprozenten für wissenschaftliche Mitarbeiter und rund 300% für studentische Hilfskräfte ausgestattet werden, bei einer vorgesehenen Projektlaufzeit von vier Jahren. Ausserdem übernahm die Projektgruppe des CHTK gewisse Vorarbeiten und Softwarekomponenten der weiter fortgeschrittenen Partnerprojekte in Berlin und Wien.[5] In einer gemeinsamen Erklärung der drei Projekte wurden die Grösse des angestrebten Korpus auf 20 Mio. Textwörter pro beteiligten Partner und Richtlinien für den Korpusaufbau – ausgewogene zeitliche Streuung über das gesamte 20. Jahrhundert und Berücksichtigung möglichst verschiedener Textsorten – festgesetzt[6] und die eingegangene Verpflichtung vertraglich geregelt. Im September 2005 stiess als letztes Projekt das *Korpus Südtirol*[7] zum Gesamtprojekt hinzu, welches die vereinbarten Vorgaben übernahm. Als Name für das Gesamtkorpus wurde *Korpus C4* gewählt.

Die zeitliche Inkongruenz auf der einen, rechtliche Fragen auf der anderen Seite, dazu eine unterschiedliche finanzielle und personelle Ausstattung sowie teils divergierende Partialinteressen haben die Gesamtzielsetzung des Projektes ein Stück weit beeinflusst. So wurde auf den Abschluss des Projekts hin nicht mehr das Ziel verfolgt, ein einziges gemeinsames Korpus aufzubauen, sondern es sollte jede Institution ihr eigenes Korpus nach gemeinsam erarbeiteten Standards entwickeln, das über ein verteiltes System in das Korpus C4 einfliessen sollte.[8] Trotz unterschiedlicher Rahmenbedingungen und zeitlicher Inkongruenz schien es bis zum Release im Februar 2009 zu gelingen, vier Teilkorpora zu einem gesamtdeutschen Korpus zu vereinen.[9]

2 Repräsentativität und Ausgewogenheit

Das Korpus C4 soll die deutsche Sprache des 20. Jahrhunderts abbilden. Es besteht aus vier gleich grossen Teilkorpora aus Deutschland, Österreich, Südtirol und der Schweiz und geht damit davon aus, dass wir es mit vier verschiedenen Varietäten des Deutschen zu tun haben, die strukturell über analoge sprachliche Ressourcen verfügen, sodass sie nicht im Verhältnis

---

[3] Internes Dokument "Erklärung der Akademien" vom 12.03.2001 zwischen Vertretern der Berlin-Brandenburgischen Akademie der Wissenschaften BBAW, der Österreichischen Akademie der Wissenschaften ÖAW und der Schweizerischen Adademie der Geistes- und Sozialwisschenschaften SAGW.

[4] Cf. Österreichische Akademie der Wissenschaften (2005–).

[5] Von Berlin die Einteilung in vier Werkkategorien à je 5 Mio. Textwörter; XML/TEI als Annotierungsstandard; Ausgestaltung der Textheader; von Wien der dort eigens entwickelte XML-Editor *Corpeduni*.

[6] Internes Dokument "Vereinbarung des AAC, des DWDS-Deutschland und des DWDS-Schweiz zum Aufbau eines gemeinsamen Korpus. Berlin, 24./25. Januar 2005".

[7] Cf. Freie Universität Bozen et al.

[8] Während Deutschland, Österreich und Südtirol aus ihren weit umfänglicheren Korpora nur eine Teilmenge für das C4-Korpus ausgewählt haben, entsprechen die 20 Mio. Textwörter in der Schweiz praktisch vollumfänglich dem Gesamtkorpus. Zu den Herausforderungen einer verteilten Korpusabfrage vgl. den Beitrag von Tobias Roth (im ersten Band).

[9] Ob es dem AAC in Wien angesichts eines Leiterwechsels und einer erneuten Evaluation möglich ist, das vereinbarte Korpus zur Abgabe freizugeben, ist derzeit leider offen.

zu ihrer Sprecherzahl, sondern als gleich grosse und damit als gleichwertige Textsammlungen vertreten sind.[10] Damit erhalten die Teilkorpora von Österreich, Südtirol und der Schweiz ein verhältnismässig grösseres Gewicht gegenüber dem bundesdeutschen Korpus. Dafür treten Unterschiede in den vier verschiedenen Varietäten umso stärker hervor. Es handelt sich dabei um eine Konzession der quantitativen Verhältnisse zugunsten der qualitativen einer plurizentrischen Sprache.[11]

Im Folgenden sollen einige methodische und praktische Vorüberlegungen zum Verhältnis eines Korpus zur in ihm enthaltenen Sprache gemacht werden.[12]

Ein Korpus repräsentiert eine sprachliche Welt bzw. es bildet sprachliche Welten in unterschiedlicher Breite oder Vollständigkeit ab. Von einem Gottfried-Keller-Korpus würde man erwarten, dass es alle bekannten Texte Kellers enthält. Es repräsentiert also die Gesamtheit der Texte Kellers. Aber wie ist es bei einem Korpus einer ganzen Epoche? Klar ist einzig: Es ist theoretisch und praktisch unmöglich, jeden schweizerischen Text des 20. Jahrhunderts zu identifizieren und ins Korpus aufzunehmen. Es braucht eine Auswahl, die nur einen winzigen Ausschnitt aus der Gesamtheit aller produzierten Texte enthält und die jene Gesamtheit in angemessener Weise repräsentiert.

Das CHTK – so eines der formulierten Projektziele – sollte einen in inhaltlicher, stilistischer, formaler und zeitlicher Hinsicht möglichst vielfältigen, ausgewogenen und repräsentativen Querschnitt schriftsprachlicher Texte des 20. Jahrhunderts in der Schweiz enthalten und neben seiner Funktion als Teilkorpus des Korpus C4 auch als Referenzkorpus der deutschen Standardsprache in der Schweiz dienen. Es stellte sich die Frage, was unter Ausgewogenheit und Repräsentativität angesichts der angestrebten Vielfalt verstanden werden sollte und wie diese hergestellt werden sollten.

Zum einen determinieren der Forschungskontext und rechtliche Rahmenbedingungen, wie ein Korpus zusammengesetzt sein kann bzw. seine Zusammensetzung hängt von der Beantwortung von folgenden pragmatischen Vorfragen ab: Welchen Einfluss üben urheberrechtliche Fragen auf den Aufbau des Korpus aus? Dürfen Texte jüngeren Publikationsdatums digitalisiert und in einem Korpus veröffentlicht werden? Wie bestimmt die Zusammenarbeit mit anderen Korpusprojekten den Korpusaufbau? Beeinflussen bestimmte Strukturen der Partnerprojekte das eigene Korpusdesign? Wenn ja: in welchem Ausmass?

Zum anderen sind methodische und theoretische Fragen zu klären: Was ist die Sprachwirklichkeit des 20. Jahrhunderts? Wodurch wird diese konstituiert? Welche Texte gehören zwingend oder im Minimum zu dieser Sprachwirklichkeit? Sind es nur gedruckte oder auch handschriftliche Texte? Sind es nur öffentlich zugängliche oder auch private Texte?

In erster Linie macht die gesprochene Sprache, mit dem grössten Anteil an der gesamten Sprachproduktion, die Sprachwirklichkeit aus. Gesprochene Sprache lässt sich wegen ihrer Flüchtigkeit und Individualität aber nicht in ihrer Gesamtheit festlegen, geschweige denn in einem Korpus qualitativ und quantitativ unproblematisch erfassen – abgesehen davon, dass sie nur mit viel Aufwand überhaupt in einer schriftlichen Form fixierbar ist.

Nimmt man aber als Hauptbasis für ein Korpus der Einfachheit halber die geschriebene Sprache, steht man u.a. auch vor der Frage: Wird Repräsentativität erreicht, indem man einen ausgewogenen Querschnitt der produzierten Texte aufnimmt, oder ist nicht viel wichtiger, welche Texte wirklich gelesen, rezipiert wurden? Die Sprachrealität – das ist in erster Linie

[10] Zum Begriff der nationalen Varietäten cf. Ammon (1995) und Kap. 3.8 unten.

[11] Eine Sprache mit mehreren, normativ gleichberechtigten Zentren oder Varietäten. Vgl. dazu grundlegend Ammon (1995).

[12] Vgl. dazu auch Biber et al. (1998).

die tatsächlich gebrauchte Sprache. Aber wie können die Bedeutung und die Wirkung von einzelnen Texten oder Texttypen im Hinblick auf die Sprachrealität festgestellt werden? Welche Texte sind in welcher Hinsicht und mit welchen Folgen wirklich wichtig? Wie sind die Textsorten qualitativ zu bestimmen und quantitativ zu gewichten?

Der bildungsbürgerliche Ansatz, wie er in den klassischen Wörterbüchern abgebildet ist, gewichtet anspruchsvolle literarische Werke weit stärker als "Massenprodukte" aus der Journalistik, der Werbung, der Trivialliteratur oder als die Vielzahl der individuellen Alltagstexte. Nähme man dagegen einzig die Produktionsmenge als Kriterium, würden die Sprachwerke von bspw. Robert Walser oder Gertrud Leutenegger, auch wenn sie kanonisiert und Schulstoff geworden wären, gegenüber dem Boulevardblatt *Blick*, gegenüber dem Meldezettel an einer Hotel-Rezeption, gegenüber amtlichen Formularen und Mitteilungen, gegenüber Bedienungsanleitungen und Flugblättern etc. weit in den Hintergrund treten. Aber selbst wenn man eine präzise Vorstellung von der Bedeutung und Wirkung einzelner Textsorten hätte und ihre Auswahl für ein Korpus danach richten könnte: Über die quantitative Verteilung von Textsorten an der Gesamtproduktion gibt es keine verlässlichen Zahlen. Der diachrone Anspruch verschärft diese Problematik: Ob die Gesamtheit der überlieferten Texte aus bspw. den 1920er Jahren der damaligen Produktion oder der nachmaligen Wirkung auf die Sprachrealität entspricht, kann nicht eruiert werden.

Ebenso gibt es keine verlässlichen Zahlen über die Geschlechterverteilung bei der Autorschaft der schriftlichen Texte, wenngleich der Bias zugunsten der Autoren offensichtlich ist. Weiter ist die Grenze zwischen Fachsprache und Alltagssprache fliessend und verändert sich über lange Zeit betrachtet. Das 20. Jahrhundert ist geprägt durch einen starken Ausbau der Fachsprachen, die durch fachexterne Kommunikation Eingang in die Alltagssprache finden. Repräsentativität im definierten statistischen Sinn kann bei einem Sprachkorpus weder theoretisch noch praktisch erreicht werden, weil die Grundgesamtheit der Daten – in unserem Fall: Die deutsche Standardsprache in der Schweiz im 20. Jahrhundert – für keinen Zeitpunkt präzise bestimmt werden kann.[13]

In der Korpuslinguistik muss die Frage der Repräsentativität eines Referenzkorpus für eine Sprache daher nicht nur theoretisch sondern auch praktisch-pragmatisch angegangen werden. Dem Begriff *Repräsentativität* haftet der Anspruch und die Problematik an, dass er die objektiv gültige Erfassung einer Sprache – ein exaktes Spiegelbild der tatsächlichen Sprachrealität – suggeriert.[14]

Aber die obige lose Aufzählung zeigt zur Genüge, dass sich Korpusprojekte im Hinblick auf Repräsentativität notwendigerweise auf dünnem Eis bewegen, bzw. eine Übersicht über die Grundgesamtheit von Sprache nicht vorgelegt werden kann, und so erstaunt es kaum, dass kein Korpusprojekt einen statistisch bestimmten, wissenschaftlich abgesicherten Weg im Umgang mit dem Problem der Repräsentativität gefunden hat. In der Regel wird die Korpusstruktur pauschal mit dem Hinweis auf die Vielfältigkeit der aufgenommenen Texte angegeben.[15] Mit anderen Worten: Repräsentativität wird zwar gerne postuliert, methodisch

---

[13] "The totality of the verbal interactions of a specific language community includes idiolects, sociolects, dialects, regional variants, languages for special purposes, eighteenth-century language and contemporary language, female language and male language, slang and jargon, and innumerable other kinds of language we can sometimes distinguish" (Teubert/Čermáková 2007: 61).

[14] Repräsentativität ist ein Begriff aus der Statistik und wird angewendet, wenn eine Grundgesamtheit, die untersucht werden soll, zu gross ist, als dass sie insgesamt erfasst oder untersucht werden könnte, weshalb ein Ausschnitt davon in Stichproben erhoben wird (*representative sample*), der typisch für alle Aspekte der Grundgesamtheit sein muss. Dazu cf. Biber (2007).

[15] Vgl. etwa das *British National Corpus* (BNP): "The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written"

umgesetzt ist sie allerdings nicht – und vor dem Hintergrund der bisherigen Ausführungen wird rasch klar, weshalb dem so ist.

Die Fiktion der Repräsentativität wird heute in der Korpuslinguistik – und in der Linguistik allgemein – zunehmend als solche erkannt und verlassen. Man grenzt sich auch sprachlich davon ab und spricht lieber von *ausgewogenen* (*balanced*) *Korpora*.[16] Das Konzept der *Ausgewogenheit* nimmt gegenüber demjenigen der *Repräsentativität* v.a. die Relativierung der theoretischen Ansprüche in Kauf, indem auf einen vollständigen und zahlenmässigen Bezug zur Grundgesamtheit verzichtet wird, weil man – bezogen auf Sprache und Sprachgebrauch – die Grundgesamtheit nicht bestimmen kann. Das Bemühen um die Erfüllung eines Kriterienkatalogs für gewichtete Vielseitigkeit[17] tritt an die Stelle einer behaupteten objektiv gültigen Gesamtschau einer Sprache. Ein solcher Kriterienkatalog dokumentiert zwar auch die Mängel und Grenzen eines Korpus, hat aber den Vorteil der objektiven Überprüfbarkeit und gibt dem Nutzer ein wichtiges Instrument bei der Interpretation der erzielten Treffer einer Korpusabfrage in die Hand. Ausgewogenheit bedeutet also, dass im Korpus vielfältige, als adäquat empfundene und in Übereinstimmung mit anderen Wirklichkeitsklassifikationen (z.B. derjenigen der Bibliotheken) stehende Ausschnitte aus der Sprachwirklichkeit abgebildet werden sollen, und dass vorgängig die Kriterien festgelegt werden, die erfüllt sein sollen, um diese "reduzierte Grundgesamtheit" – ein Modell einer Grundgesamtheit – in ihrer Typizität möglichst gut zu vertreten.

3 Korpusdesign und Korpusaufbau des CHTK

Nach diesen grundsätzlichen methodischen und methodologischen Erwägungen wird im Folgenden ausführlich auf das eigentliche Design des CHTK eingegangen. Transparenz stellt im Zusammenhang mit dem Korpusdesign eines der Qualitätsmerkmale eines Korpus dar und ist für dessen Brauchbarkeit unabdingbar. Ein Korpusdesign wird durch die Grösse, die Zusammensetzung und die technische Realisierung eines Korpusprojekts in Abhängigkeit von den vorhandenen Ressourcen und dem angestrebten Zielprodukt bestimmt. Bereits vor Projektbeginn gab es vordefinierte und nicht vordefinierte Charakteristika für dieses Korpus. Vordefiniert waren, wie bereits ausgeführt, die generelle Art (Referenzkorpus der deutschen Sprache des 20. Jahrhunderts in der Schweiz) und der Umfang (20 Mio. Textwörter) des Korpus, die finanzielle und personelle Ausstattung des Projekts sowie das Schweizerische Urheberrecht, zu welchem es eine Umgangspraxis zu finden galt. Zu den nicht vordefinierten Charakteristika gehörten die Aufteilung der hundert abzudeckenden Jahre in Rechercheeinheiten, die regionale Verteilung der Texte, die quantitative Berücksichtigung der Geschlechter, die inhaltliche Verteilung der Texte auf Domänen bzw. Sachgruppen, die Verteilung der Texte auf Kriterien wie *publiziert – unpubliziert*, *formell – informell*, *Allgemeinwortschatz – Fachwortschatz* etc. Ausserdem natürlich technische Fragen der Digitalisierung, auf welche hier aber nicht weiter eingegangen wird. Im Folgenden werden zunächst vertiefende Erläuterungen zu den Faktoren Grösse, Zentralrepertoire und Urheberrecht, anschliessend zu den 'harten' Hauptkriterien Werkkategorie, Jahrhundertviertel und Sachgruppe und zu den 'weichen' Kriterien Geschlecht und regionale Verteilung gegeben.

(www.natcorp.ox.ac.uk/ Stand: 28.02.2009). Auch auf der Homepage des "Czech National Corpus CNC" findet man nur vage Angaben über die intendierte Zusammensetzung des Korpus: "CNC presents a very large, modern and valuable language and informational base" (http://ucnk.ff.cuni.cz/; Stand: 28.2.2009). In anderen Metatexten, hier auf der Seite von elsenet wird aber auch von Repräsentativität gesprochen: "It includes, by 2002, a 100 million representative corpus of written contemporary Czech" (www.elsnet.org/orgs/3017.html; Stand: 28.02.2009).

[16] Cf. Lemnitzer/Zinsmeister (2006: 50f.); Sinclair (1998: 123f.); Teubert/Čermáková (2007: 59).

[17] Die Vielseitigkeit berücksichtigt die unterschiedlichsten Bedingungen, Themen und Wirkungen von Texten, während die Gewichtung ihre Bedeutung im "Gesamtkonzert" einbezieht.

Der Kriterienkatalog hat sich im Verlauf des ersten Projektjahres verfestigt und verfeinert, wurde, wo immer möglich, streng eingehalten und liegt nun dem CHTK als Struktur zugrunde. Die drei Hauptkriterien berücksichtigen die Art der Texte, ihr Erscheinungsdatum und den Inhalt. Sie sind grundsätzlich gleich gewichtet, aus praktischen Gründen wurden sie aber hierarchisiert nach der Reihenfolge obiger Aufzählung: Jede Werkkategorie wurde in Jahrhundertviertel aufgeteilt, dieses wiederum jeweils auf die Sachgruppen.[18]

3.1 Grösse

Der Faktor Grösse eines Korpus sollte – berücksichtigt man die Entwicklungen der vergangenen Jahrzehnte – sinnvollerweise nicht mehr das wichtigste Merkmal und Qualitätskriterium eines Korpus bilden. Nachdem das Pionierkorpus *Brown Corpus* 1967 mit 1 Mio. Textwörtern für damalige Verhältnisse alle Grenzen sprengte, erreichte *The Birmingham Collection of English Text* 1985 bereits 20 Mio., *The Bank of English* Mitte der 90er Jahre 320 Mio.[19], während heutige Grosskorpora wie das *COSMAS* am IDS in Mannheim laut eigenen Angaben in 65 Subkorpora 3,3 Milliarden Textwörter zur Abfrage bereitstellen[20] – ganz abgesehen von Suchmaschinen.

wie Google, welche bestrebt sind, möglichst die Gesamtheit des World Wide Web zu indexieren. Nicht die absolute Grösse eines Korpus ist also für seine Bewertung zentral, sondern ihr Verhältnis zur im Korpus abgebildeten Sprache oder Sprachvarietät und vor allem zum berücksichtigten Datenmaterial. Korpora, welche opportunistisch möglichst viele digital verfügbare Texte sammeln oder solche, welche grosse Mengen an Texten ohne differenzierte Annotation bereitstellen, können mit ganz anderen Gesamtgrössen operieren als Korpora, die für einen bestimmten Zweck eine bestimmte Struktur und Bearbeitung erfordern. Das CHTK ist ein vergleichsweise sehr kleines Korpus jedoch mit hohem Anspruch der Ausgewogenheit der aufzunehmenden Texte. Die damit verbundene aufwändige Recherchier- und Digitalisierungsarbeit bei schwer zu beschaffenden Texten sowie die aufgewendete Sorgfalt bei der Annotierung auch sperriger Texte lassen die 20 Mio. in einem anderen Licht erscheinen. Die Faktoren Quantität und Qualität eines Korpus stehen in einem gegenseitigen Abhängigkeitsverhältnis und sind unterschiedlich zu beurteilen, je nachdem, wofür ein Korpus dienen soll. So sind Frequenzanalysen, welche die Entwicklung der deutschen Sprache in der Schweiz allgemein oder in Einzelwort- oder Phraseologismenanalysen zuverlässig abbilden sollen, mit dem CHTK zwar nicht sinnvoll. Dafür ist die durchschnitt-liche Zahl von 200'000 Wörtern pro Jahr zu gering.[21] Der vorgesehene Zweck des CHTK, als Subkorpus des *Korpus C4* Teil der Basis für ein Digitales Wörterbuch der Deutschen Sprache des 20. Jahrhunderts, mithin ein zuverlässiges Instrument für die Lexikographie zu sein, kann, im Sinne des folgenden Zitats von Sinclair, dagegen erfüllt werden:

> Ein *Referenzkorpus* ist ein Korpus, das konzipiert worden ist, um umfassende Informationen über eine Sprache zu liefern. Es hat den Anspruch, groß genug zu sein, um alle wichtigen Varietäten einer Sprache einschließlich des jeweiligen Wortschatzes zu repräsentieren, so daß es als Grundlage für verläßliche Grammatiken, Wörterbücher, Thesauri und andere Nachschlagewerke über Sprache dienen kann. (Sinclair 1998: 123; Hervorh. im Orig.)

---

[18] Damit tritt das bisher linguistisch verwendete, jedoch nicht definierte Konzept der *Textsorte* in den Hintergrund. Da jedoch keine abschliessende, allgemein akzeptierte Textsortenklassifikation vorliegt, lässt sich das Konzept in der Praxis höchstens als heuristisches Prinzip einsetzen.

[19] Cf. Sinclair (1998: 112).

[20] www.ids-mannheim.de/cosmas2/uebersicht.html (Stand 17.2.2009).

[21] Teubert/Čermáková (2007) stellen diese Art statistischer Repräsentativität für jedes Referenzkorpus grundsätzlich in Abrede: "The presence or absence of words less frequent is as unpredictable as the winning numbers in a lottery. But even if we only consider the most frequent part of the vocabulary we find ourselves at a loss" (Teubert/Čermáková 2007: 64).

Erste Probeauswertungen sowie Vergleichsauswertungen mit anderen Korpora zeigen auch, dass ein kleines Korpus bei vielen Fragestellungen durchaus mithalten kann.[22]

3.2 Zentralrepertoire

Das CHTK soll das Zentralrepertoire der deutschen Sprache in der Schweiz repräsentieren. Mit dem Begriff *Zentralrepertoire* meinen wir denjenigen Bereich des Wortschatzes, über welchen eine ausgebildete und durchschnittlich gebildete Person zumindest passiv verfügen kann und mit welchem alle Leserinnen und Leser bzw. Hörerinnen und Hörer über entsprechende Texte potentiell in Berührung kommen können (vgl. Schnörch 2002). Das Zentralrepertoire ist also mehr als der *Allgemeinwortschatz* und deutlich mehr als der *Grundwortschatz*, der eher den durchschnittlich aktiv beherrschten Wortschatz meint.[23]

Der Begriff *Zentralrepertoire* gewinnt aber an Kontur, wenn man ihn als Gegensatz zu *Fachsprache* versteht, also gegen den Bereich abgrenzt, der nur durch berufliche oder freizeitliche inhaltliche Spezialisierung erschlossen werden kann. Bei der Fachkommunikation unterscheidet man fachinterne und fachexterne Kommunikation.[24] Je nachdem, ob ein Gärtner über Fachangelegenheiten mit seinem Kollegen oder mit einem Kunden spricht, verwendet er ein anderes Repertoire. Dasselbe gilt für die Anwältin, die mit ihrer Sekretärin bzw. mit ihrer Kundin oder ihren Freundinnen über juristische Angelegenheiten spricht und dabei unterschiedliche fachinterne und fachexterne Repertoires verwendet.

Mit Zentralrepertoire ist die grösstmögliche Schnittmenge der verschiedenen Repertoires im allgemeinsprachlichen und fachexternen Bereich gemeint. Praktisch bedeutet diese Definition einerseits, dass spezialisierte wissenschaftliche Bücher und Zeitschriftenartikel nicht aufgenommen wurden, populäre Fachzeitschriften dagegen schon. Zum anderen resultierte aus diesem Verständnis der im Korpus abzubildenden Sprache das Bemühen um kurze Texte: Um Überrepräsentationen individueller Stile, thematischer Fokussierung oder einzelner Jahre zu verhindern, wurden für jedes Kriterienbündel (Sachgruppe pro Jahrhundertviertel pro Werkkategorie) mehrere Texte aufgenommen.[25]

3.3 Urheberrechte

Ein Punkt, der in vielen Korpusprojekten eher summarisch beschrieben wird, ist der Umgang mit dem Urheberrecht.[26] Dies hat einen guten Grund: Bei der Ausgestaltung des Urheberrechts durch die Politik hatte man nicht die wissenschaftlichen Korpusprojekte im Blick, sondern den Buch- und Pressemarkt bzw. den adäquaten Ausgleich der ökonomischen Interessen, die sich auf der Basis des Schreibens und des gesellschaftlichen Verteilens von Schreibprodukten ergeben.[27] Die Anwendung des Urheberrechts auf ein Korpusprojekt kann sowohl praktisch undurchführbar wie auch unbezahlbar sein. Es war mit Sicherheit nicht die Meinung des Gesetzgebers, Forschung ohne finanzielle Interessen im Textbereich verhindern zu wollen.

Aus teilweise ökonomischen, teilweise nationalistischen oder ideologischen und teilweise auch ideellen Gründen wie bei der Produktion von Druckerzeugnissen sind im Bereich der

---

[22] Cf. auch aus der Kindersprachforschung Rowland et al. (2008).

[23] Wobei diese Begriffsabgrenzung hier nicht methodisch verankert ist – auch Sinclair (1998: 123), der den Begriff gebraucht, definiert ihn nicht präziser.

[24] Cf. Gläser (1990) oder Niederhauser (1999).

[25] Vgl. unten die Statistik in Kap. 3.5.

[26] Vgl. jedoch Uwe Quasthoffs Angaben zum Projekt Deutscher Wortschatz in diesem Heft. Das Leipziger Projekt Deutscher Wortschatz sammelt Textdaten in der Form von Sätzen.

[27] Einen Einblick in die gegenwärtig laufende Revision des Urheberrechts gibt die folgende Website: www.urheberrecht.ch.

Textdigitalisierung gewichtige Akteure wie Google, National- und Universitätsbibliotheken aufgetreten, die ein Interesse daran haben, noch wesentlich grössere Textmengen als bisher vorhanden oder angestrebt zu digitalisieren. Dadurch ist in der Gesellschaft ein Bewusstsein dafür entstanden, dass im Interesse der Forschung und der Informationsbeschaffung ein neuer Ausgleich zwischen den berechtigten Interessen der Urheber an ihrem Werk und den Ansprüchen der Gesellschaft auf Informations- und Wissenserschliessung gefunden werden muss.

Im CHTK ist ein Teil der aufgenommenen Texte urheberrechtlich geschützt. Ein Korpus, das auch die Gegenwartssprache berücksichtigt, besteht zwingend auch aus solchen Texten. Nur Texte von Autoren, die vor dem 1. Juli 1943 verstorben sind, sind frei.[28] Eine Beschränkung auf diese Texte war nicht denkbar. Ebenfalls nicht in Betracht kam, unbekannte Urheber ausfindig machen zu wollen[29], mit allen Urhebern in Kontakt zu treten und Verhandlungen zu führen. Bei einem Korpus wie dem CHTK, das aus einigen Tausend Texten besteht, wäre ein solches Vorhaben schlicht unmöglich. Eine globale Lösung zusammen mit der Verwertungsgesellschaft Pro Litteris kam ebenfalls nicht zustande – nicht zuletzt, weil Pro Litteris weder Alltagstextautoren noch unbekannte Autoren vertritt.

Es wurde deshalb ein Weg gesucht, der die Ausgewogenheit des Korpus garantiert, dabei aber keine zu grossen Risiken für das Projekt eingeht und die Rechteinhaber nicht brüskiert. Dieser Weg sieht so aus, dass als Suchresultat jeweils nur ein Zitat angezeigt wird – ähnlich wie das in Wörterbüchern üblich ist. Die Rekonstruktion eines integralen Textes aus den Suchresultaten ist damit unmöglich. Zusätzlich wurde von noch im Handel befindlichen Texten immer nur ein Ausschnitt digitalisiert. Und schliesslich wurde und wird Rechteinhabern angeboten, dass sie ihre Textausschnitte auf expliziten Wunsch aus dem Korpus entfernen lassen können. Dies dürfte aber wohl eher selten der Fall sein, da die Aufnahme eines Textes in das CHTK dessen Bedeutung und Wahrnehmung in der Öffentlichkeit erhöht, ohne die kommerziellen Chancen zu beeinträchtigen. Beanstandungen sind bisher nicht eingetroffen.

3.4 Werkkategorie

In formaler Hinsicht ist das CHTK in die vier Kategorien Belletristik, Sachtexte, Journalistische Prosa und Gebrauchstexte eingeteilt. Die vier Kategorien sind zu gleichen Teilen im Korpus vertreten und stellen je einen Viertel des vorgesehenen Textwörterbestandes, d.h. rund 5 Millionen Textwörter. Die Einteilung der Texte in diese vier Werkkategorien basiert nicht auf einer exakten Texttypologie. Die Unterscheidung diente primär als heuristisches Prinzip für den Korpusaufbau und ist deshalb praktisch-pragmatisch motiviert. Auf einen mehrere Dimensionen der Textsortenbeschreibung umfassenden Kriterienkatalog wurde zugunsten einer intuitiv verständlichen Werkkategorienunterscheidung verzichtet. Je differenzierter nämlich eine Matrix der Textsortenbeschreibung ausgearbeitet wird, desto mehr Eigenschaften der Matrix haben keine vergleichbare Entsprechung in der Textwirklichkeit, es entstehen Löcher und Überdeterminationen, die darauf zurückgehen, dass die Welt der Texte historisch und nicht entlang den Leitlinien eines Modells entstanden ist. Ebenso beruht die Viertelung der Anteile nicht auf einer quantitativen Analyse der vier Kategorien, sondern auf dem Willen, sie zu gleichen Anteilen im Korpus vertreten zu haben.[30]

---

[28] Das heute geltende, jedoch in Revision befindliche Urheberrechtsgesetz, das den Urheberschutz von 50 auf 70 Jahre nach dem Tod erhöhte, ist am 1. Juli 1993 in Kraft getreten. Da die Schutzfrist nicht rückwirkend erhöht wurde, gilt bis 2013 der 1. Juli 1943 als Stichdatum. Das Gesetz kann unter dem folgenden Link eingesehen werden: www.admin.ch/ch/d/sr/231_1/ (Stand 28.2.2009).

[29] Bei über der Hälfte der Texte des CHTK ist die Autorschaft unbekannt; vgl. dazu auch Anm. 46 und Tabelle 4.

[30] Ausserdem wurden, wie bereits erwähnt, diese Kategorien und ihre Anteile am Gesamt des Korpus aus Kompatibilitätsgründen vom Partnerprojekt DWDS übernommen; zu den definitiven Zahlen jeder Kategorie

Gebrauchstexte sind im CHTK definiert als Texte, die primär "für jemanden" geschrieben sind und einen begrenzten Kreis von Adressaten anvisieren, die in einer konkreten Situation Hilfe, Anweisung oder Information etc. benötigen (erweiterte appellative Funktion). Typische Beispiele sind etwa Gebrauchsanweisungen, Werbung, Ratgeberliteratur oder Kochrezepte. Neben gedruckten und publizierten Texten beinhaltet das CHTK auch Gebrauchstexte, die schwer zugänglich sind. Sinnbildlich dafür steht eine virtuelle Waschküchenordnung von 1920, ein Text, der natürlich nicht konkret und gezielt gesucht werden kann, sondern als Typus nur über aufwändige Recherche in Archiven, allenfalls in Antiquariaten und Brockenhäusern auffindbar ist. Dank der Zusammenarbeit mit dem Staatsarchiv Aarau, dem Sozialarchiv Zürich, dem Schweizerischen Wirtschaftsarchiv und dem Sportmuseum Schweiz konnten solche und andere Texte wie private Briefe, Rechnungen, Fragebögen, Rats- oder Vereinsprotokolle etc. über die gesamte abzudeckende Zeitspanne gefunden, digitalisiert und ins Korpus aufgenommen werden. Die Berücksichtigung dieser sonst in Sprachkorpora im Allgemeinen nicht enthaltenen Texte zeichnet das CHTK in besonderem Masse aus und verleiht der Ambition 'ausgewogen' eine weitere neue Dimension.[31]

Sachtexte sind im CHTK definiert als Texte, in denen primär "über etwas" geschrieben wird (referentielle Funktion). Dazu gehören beispielsweise Ausschnitte aus populärwissenschaftlichen Sachbüchern, Artikel in Fachmagazinen, die in Kiosken verkauft werden, aber auch Biographien oder etwa eine Abhandlung über die Trachten im Emmental. Texte, wie z.B. Dissertationen, die auf fachinterne Kommunikation ausgerichtet sind und einen spezifischen Fachwortschatz voraussetzen müssen, repräsentieren nicht das Zentralrepertoire und wurden nicht ins Korpus aufgenommen.

Zur Werkkategorie Belletristik gehören vor allem kürzere literarische Texte von Deutschschweizer Autorinnen und Autoren aus allen Gattungen mit Ausnahme der Lyrik – also Romane, Erzählungen, Theaterstücke und dergleichen. Über die Aufnahme ins CHTK entschied nicht in erster Linie die Zugehörigkeit zum Kanon der Deutschschweizer Literatur, sondern das Publikationsjahr und der Urheberrechtsstatus des Textes, sowie das Geschlecht der Autorin bzw. des Autors.

Um eine möglichst grosse Bandbreite verschiedener sprachlicher "Handschriften" abzudecken, wurden auch hier vor allem kleinere Texte möglichst verschiedener Provenienz ins Korpus aufgenommen. Lyrik hingegen wurde nicht berücksichtigt, da sie in der Regel davon lebt, experimentell mit der Sprache umzugehen und deshalb einem speziellen Repertoire angehören kann, das über das Zentralrepertoire hinausgeht.

Zur Werkkategorie journalistische Prosa gehören Artikel aus der Deutschschweizer Tages- und Wochenpresse sowie aus Fachperiodika, sofern sie sich als fachexterne Kommunikation an ein grösseres Publikum richten. Bei der Beschaffung von journalistischer Prosa wurden über 300 verschiedene Schweizer Zeitungen und Zeitschriften aus der gesamten Deutschschweiz berücksichtigt.

Die Abgrenzung zwischen Artikeln aus Fachzeitschriften und Texten aus Sachbüchern ist natürlich nicht ganz unproblematisch. Die Schwierigkeit liegt darin, dass Sachtexte primär ein eher inhaltlich-funktionales Kriterium, journalistische Prosa dagegen ein formales Kriterium ist und es deshalb zu Fällen von Kreuzklassifikationen kommt, in denen ein Text beide Kriterien erfüllt. In der Praxis wurde versucht, die Klassifikation aufgrund der Publikations-

---

vgl. Tabelle 2. Eine analoge Festsetzung wurde auch bezüglich des Geschlechterverhältnisses der TextautorInnen getroffen. Allerdings hatte diese letzte Festsetzung mit der Realität der zahlenmässigen Dominanz männlicher Textproduzenten zu kämpfen ebenso wie mit der Unbestimmbarkeit der Autorschaft in etwa der Hälfte der aufgenommenen Texte.

[31] Nicht berücksichtigt werden konnten allerdings handschriftliche Texte, da deren Digitalisierung zu aufwendig ist.

form durchzuführen: im normalen Detailhandel erhältliche Periodika gehören eher zur journalistischen Prosa, während einmalige Publikationen zu spezifischen Themen eher zu den Sachtexten gezählt werden.

Bei der Beschaffung der journalistischen Texte wurden zuerst ganze Zeitungen bzw. Zeitschriften integral aufgenommen. Damit war aber ein relativ hohes Mass an Redundanz, v.a. bei Werbungen und Anzeigen, verbunden, und umgekehrt waren bestimmte Zeiträume und/oder Sachgebiete unterrepräsentiert. Ausserdem gestaltete sich die Digitalisierung einer ganzen Zeitung derartig aufwendig, dass mit den vorhandenen Mitteln nur wenige Exemplare bewältigt worden wären. Aus diesen Gründen wurde die Strategie im Laufe der Zeit modifiziert: Wie bei den Gebrauchstexten wurden Zeitungsartikel aus Sammlungen in Archiven gezielt nach den Korpuskriterien zusammengetragen und einzeln ins Korpus aufgenommen.

3.5 Sachgruppen

Um eine breite inhaltliche Streuung der Korpustexte zu erreichen, wurde nach bestehenden Sachklassifikationssystemen gesucht, die gewährleisten sollen, dass einerseits kein wichtiger sachlicher und damit natürlich auch sprachlicher Bereich ausgelassen würde, und die andererseits mit anderen Korpora oder auch Bibliotheken wenigstens minimal kompatibel wären. Nach Möglichkeit sollte sich das Korpus an einen für sprachliche Zwecke adäquaten Standard anschliessen. Zu diesem Zweck wurden zwei verschiedene Klassifikationssysteme evaluiert: Die Dewey Decimal Classification (DDC)[32] und die Schlagwortnormdatei (SWD)[33]. Diese Systeme werden von zahlreichen europäischen Bibliotheken und Bibliotheksverbünden eingesetzt, um ihre Bestände nach inhaltlichen Kriterien zu klassifizieren und zu beschlagworten. Als für unsere Zwecke gut brauchbar hat sich das SWD herausgestellt, das u.a. auch von der Schweizerischen Landesbibliothek benutzt wird. Es besteht aus 36 inhaltlichen Oberkategorien (Sachgruppen). Diese fächern sich differenziert auf in verschiedene Ebenen von Unterkategorien. Als Beispiel sei hier die Sachgruppe 3 *Religion* aufgeführt[34]:

3 RELIGION
3.1 ALLGEMEINE UND VERGLEICHENDE RELIGIONSWISSENSCHAFT, NICHTCHRISTLICHE RELIGIONEN
3.1p Personen
3.2 BIBEL
3.2a Altes Testament
3.2aa Teile des Alten Testaments
3.2b Neues Testament
3.2ba Teile des Neuen Testaments
3.2p Personen
3.3 KIRCHENGESCHICHTE
Dogmen- und Theologiegeschichte → auch 3.4a – 3.4b
Frömmigkeitsgeschichte → auch 3.5a
Missionsgeschichte → auch 3.5cb
3.3a Antike
3.3b Mittelalter
3.3c Neuzeit
Personen → 3.6p
3.4 SYSTEMATISCHE THEOLOGIE
3.4a Allgemeines, Fundamentaltheologie
3.4b Dogmatik
Theologische Anthropologie → 3.4c

[32] www.oclc.org/dewey/ (Stand: 28.2.2009).

[33] www.d-nb.de/standardisierung/normdateien/swd.htm (Stand: 28.2.2009).

[34] Zitiert nach der Homepage der deutschen Nationalbibliothek, wo das System weiterentwickelt und betreut wird: www.d-nb.de/standardisierung/pdf/swd_syst.pdf (Stand: 28.2.2009).

3.4c Theologische Anthropologie, Christliche Ethik
Personen → 3.6p
3.5 PRAKTISCHE THEOLOGIE
3.5a Liturgik, Frömmigkeit
3.5b Homiletik, Katechetik
3.5ba Homiletik
3.5bb Katechetik, Christliche Erziehung, Kirchliche Bildungsarbeit
3.5c Seelsorge, Mission
3.5ca Seelsorge
3.5cb Mission, Kirchliche Sozialarbeit
Personen → 3.6p
3.6 KIRCHE UND KONFESSION
Kirchenrecht → 7.13
3.6a Katholische Kirche
3.6b Evangelische Kirchen
3.6c Ostkirchen und andere christliche Religionsgemeinschaften und Sekten
3.6p Personen zu 3.3 – 3.6

Im Hinblick auf die 20 Mio. angestrebten Textwörter ist eine solche hochdifferenzierte Unterkategorisierung nicht sinnvoll, weshalb die aufgenommenen Texte nach den Oberbegriffen jeder Sachgruppe klassifiziert wurden.[35] Ausserdem wurden zwei Sachgruppen, die zu allgemein und inhaltlich zu unspezifisch waren, von Anfang an weggelassen bzw. in anderen Sachgruppen subsumiert.[36]

Die definitive Liste, nach der die 20 Mio. Textwörter des CHTK inhaltlich klassifiziert wurden, sieht folgendermassen aus:

2. Schrift, Buch, Presse
3. Religion
4. Philosophie
5. Psychologie, Esoterik
6. Kultur, Erziehung, Bildung, Wissenschaft
7. Recht, allgemeine Verwaltung
8. Politik, Militär
9. Soziologie, Gesellschaft, Arbeit, Sozialgeschichte
10. Wirtschaft, Verkehr, Umweltschutz, Raumordnung
11. Sprache
12. Literatur
13. Bildende Kunst, Photographie
14. Musik
15. Theater, Tanz, Film, Rundfunk
16. Geschichte
17. Volkskunde, Völkerkunde
19. Geowissenschaften
20. Astronomie, Weltraumforschung
21. Physik
22. Chemie
23. Allgemeine Biologie, Mikrobiologie
24. Botanik
25. Zoologie
26. Anthropologie
27. Medizin
28. Mathematik
29. Stochastik, Operations Research
30. Informatik, Datenverarbeitung
31. Technik

[35] Eine spätere Verfeinerung der Klassifikation bleibt jedoch theoretisch möglich.

[36] Sachgruppe 1 "Allgemeines, Interdisziplinäre Allgemeinwörter" und Sachgruppe 18 "Natur, Naturwissenschaften allgemein". Die beiden Sachgruppen fehlen entsprechend in der nachfolgenden Auflistung.

32. Landwirtschaft, Garten
33. Hauswirtschaft, Körperpflege, Mode, Kleidung
34. Sport
35. Spiel, Unterhaltung
36. Basteln, Handarbeiten, Heimwerken

Die Gruppe 12 "Literatur" umfasst nicht die Primärliteratur, wie sie durch die Werkkategorie Belletristik abgedeckt wird. Diese, die eigentlichen literarischen Erzeugnisse, entziehen sich in aller Regel der inhaltlichen Klassifizierung, weshalb sie im Textkorpus auch keiner Sachgruppe zugeordnet wurden. Sachgruppe 12 umfasst nur das sekundäre Schrifttum zur Literatur und zum Literaturbetrieb. Nach Sachgruppen klassifiziert sind also Gebrauchstexte, Sachtexte und journalistische Texte.

Diese 36 bzw. von uns verwendeten 34 inhaltlichen Kategorien des SWD konnten jedoch nicht nach quantitativen Kriterien zu gleichen Anteilen vergeben werden. Die oben methodisch behandelten Fragen nach der ausgewogenen Abbildung der Sprachrealität in einem Korpus und nach der Frage, ob die produzierte oder die rezeptierte Textmenge als "reduzierte Grundgesamtheit" des Korpus angesehen werden soll, verschärft sich an diesem Punkt. Würde jede Sachgruppe quantitativ gleich gewichtet, ergäbe dies ein massives Übergewicht bspw. der 'schöngeistigen Künste' (Gruppen 11–14), bei ebenso massiver Untergewichtung von so umfangreichen und ungleich konstituierten Gruppen wie Politik, Militär (8), Technik (31) oder Hauswirtschaft, Körperpflege, Mode, Kleidung (33). Wer würde akzeptieren, dass Gruppe 20. "Astronomie, Weltraumforschung" ein Sechsunddreissigstel des gesamten deutschsprachigen Schrifttums in der Schweiz im 20. Jahrhundert ausmacht?

Aus diesem Grund wurde für das SCHWEIZER TEXT KORPUS eine eigene Textmengen-Hierarchie der einzelnen Sachgruppen festgelegt. In drei verschiedenen Ratingprozessen wurde versucht, intuitiv abzuschätzen, welchen prozentualen Anteil eine Sachgruppe an der gesamten Textproduktion etwa besitzt, um dann jeder Sachgruppe ein quantitatives Gewicht innerhalb des Korpus zuzuteilen. Diese Untersuchung wurde mit je einer grösseren Gruppe von Naturwissenschaftern der ETH Zürich, von Juristen der Universität Basel und von Geisteswissenschaftlern der Universität Basel durchgeführt.[37] Die Ergebnisse dieser Umfrage, in welcher nach der Gewichtung der 34 Sachgruppen gefragt wurde, flossen direkt in die Textrecherche für das CHTK ein.

Bei der Erfüllung von Kriterien waren noch andere Klippen zu umschiffen. So gibt es Sachgruppen, welche im Verlauf des Abdeckungszeitraums, also des 20. Jahrhunderts, ihre Bedeutung massiv verändert haben, bspw. Gruppe 30 "Informatik, Datenverarbeitung". Auch hier ist letztlich zu konstatieren, dass statistisch erhärtete Zahlen über die tatsächlichen Verhältnisse nicht zu eruieren sind. Neben den festgelegten und erarbeiteten Kriterien und Gewichtungen hat das Kriterium Verfügbarkeit von Texten in der Praxis den Korpusaufbau mitbestimmt.

[37] Evaluation durch die studentische Mitarbeiterin Eftychia Fountoulakis.

	1. Jahrhundertviertel		2. Jahrhundertviertel		3. Jahrhundertviertel		4. Jahrhundertviertel	
	Textwörter	Werke	Textwörter	Werke	Textwörter	Werke	Textwörter	Werke
2. Schrift, Buch, Presse	128'000	139	107'000	197	133'000	107	156'000	568
3. Religion	105'000	82	130'000	95	131'000	104	113'000	107
4. Philosophie	84'000	26	66'000	35	92'000	43	258'000	95
5. Psychologie, Esoterik	44'000	17	91'000	41	133'000	46	57'000	57
6. Kultur, Erziehung, Bildung, Wissenschaft	158'000	163	133'000	102	150'000	217	165'000	242
7. Recht, allgemeine Verwaltung	172'000	108	208'000	125	143'000	102	180'000	133
8. Politik, Militär	78'000	196	300'000	221	281'000	65	140'000	135
9. Soziologie, Gesellschaft, Arbeit, Sozialgeschichte	141'000	256	132'000	177	167'000	79	239'000	205
10. Wirtschaft, Verkehr, Umweltschutz, Raumordnung	165'000	236	191'000	428	336'000	989	143'000	224
11. Sprache	67'000	18	193'000	35	75'000	72	128'000	100
12. Literatur	68'000	71	98'000	86	69'000	88	117'000	134
13. Bildende Kunst, Photographie	74'000	43	64'000	62	144'000	71	112'000	87
14. Musik	45'000	54	93'000	93	81'000	64	136'000	183
15. Theater, Tanz, Film, Rundfunk	107'000	126	111'000	217	108'000	168	98'000	118
16. Geschichte	108'000	65	327'000	76	141'000	59	115'000	60
17. Volkskunde, Völkerkunde	92'000	20	79'000	122	134'000	42	98'000	101
19. Geowissenschaften	131'000	35	78'000	90	95'000	63	113'000	71
20. Astronomie, Weltraumforschung	71'000	19	59'000	8	54'000	54	64'000	87
21. Physik	130'000	10	80'000	29	46'000	22	62'000	20
22. Chemie	66'000	33	85'000	8	98'000	41	81'000	31
23. Allgemeine Biologie, Mikrobiologie	45'000	7	103'000	25	110'000	30	67'000	95
24. Botanik	39'000	9	60'000	20	63'000	7	83'000	6
25. Zoologie	63'000	19	117'000	73	134'000	70	100'000	106
26. Anthropologie	60'000	28	88'000	85	121'000	64	88'000	61
27. Medizin	99'000	90	124'000	113	214'000	114	194'000	133
28. Mathematik	68'000	15	69'000	15	58'000	25	69'000	21
29. Stochastik, Operations Research	38'000	25	72'000	29	62'000	63	99'000	87
30. Informatik, Datenverarbeitung	54'000	11	36'000	7	95'000	41	127'000	79
31. Technik	106'000	93	129'000	135	136'000	66	176'000	155
32. Landwirtschaft, Garten	84'000	33	138'000	90	158'000	66	140'000	99
33. Hauswirtschaft, Körperpflege	101'000	335	176'000	335	133'000	91	91'000	201
34. Sport	83'000	33	259'000	206	132'000	88	265'000	181
35. Spiel, Unterhaltung	55'000	35	79'000	77	75'000	71	79'000	150
36. Basteln, Handarbeiten, Heimwerken	30'000	14	73'000	61	54'000	53	93'000	92

Tabelle 1: Inhaltliche Zusammensetzung der im CHTK zur Verfügung gestellten Texte[38]

[38] Stand: CHTK 18.2.2009 (Zahlen gerundet).

3.6 Jahrhundertviertel

Um auch in diachroner Hinsicht ein möglichst ausgewogenes Korpus zu erhalten, wurden für jede Werkkategorie und jede Sachgruppe aus allen Jahrhundertvierteln tendenziell gleich viele Textwörter ins Korpus aufgenommen. Ein feineres Raster als die Aufteilung in Jahrhundertviertel (etwa Dekaden) war zu Projektbeginn ein Desiderat und wäre aus der Perspektive der historischen Sprachforschung wünschenswert gewesen, hätte allerdings einen die Möglichkeiten des Projekts sprengenden Mehraufwand mit sich gebracht. Die Anzahl zu beschaffender Textwörter pro Werkkategorie, Sachgruppe und Dekade wäre erheblich geringer gewesen; d.h., es hätten noch kleinere Texte in viel grösserer Anzahl beschafft werden müssen, was nicht nur den Rechercheaufwand potenziert, sondern auch den Digitalisierungsprozess verlangsamt hätte. Die Einteilung in Jahrhundertviertel ist insofern etwas problematisch, als die gleichmässige Verteilung auf Jahre und Jahrzehnte nur bedingt gewährleistet ist, bzw. umgekehrt das Klumpenrisiko für einzelne Jahre oder Jahrzehnte nicht automatisch ausgeschlossen wird.

	Werke/ Textwörter 1900–1924		**Werke/ Textwörter 1925–1950**		**Werke/ Textwörter 1950–1974**		**Werke/ Textwörter 1975–1999**		**gesamt**	
Gebrauchstexte	1'028	1'106'896	1'448	1'260'021	952	1'201'020	1'395	1'088'953	4'823	4'656'890
Sachtexte	156	1'353'146	338	1'930'650	797	1'975'894	264	2'073'653	1'555	7'333'343
Belletristik	160	1'068'685	35	1'115'080	124	984'235	47	1'007'049	366	4'175'049
Journalistische Prosa	811	498'463	1'083	1'011'682	954	977'886	1'794	1'087'926	4'642	3'575'957
gesamt	2'155	4'027'190	2'904	5'317'433	2'827	5'139'035	3'500	5'257'581	11'386	19'741'239

Tabelle 2: Texte und Textwörter jeder Kategorie pro Jahrhundertviertel[39]

3.7 Geschlecht

Damit das Korpus auch über genderspezifische Fragestellungen Auskunft geben kann, sollte bei der Beschaffung der Texte auch auf eine angemessene Repräsentation von Autorinnen und Autoren geachtet werden. Auf die Einführung einer "Quotenregelung", d.h. eine klar definierte Verteilung von Autoren und Autorinnen, musste nach langen Diskussionen und Recherchen allerdings verzichtet werden.[40] Im Verlaufe der Projektarbeit wurde festgestellt, dass Autorinnen insgesamt untervertreten sind. In der Praxis wurde deshalb bei der Beschaffung neuer Texte jeweils darauf geachtet, möglichst nur noch Autorinnen zu berücksichtigen. Dennoch konnte die Unterrepräsentation der Autorinnen nur gemildert werden. Zudem sind mit grösster Wahrscheinlichkeit auch bei den nicht-identifizierten AutorInnen diejenigen männlichen Geschlechts viel häufiger vertreten.[41]

[39] Stand: CHTK 28.2.2009, entspricht dem vermutlichen Schlussbestand Ende März 2009.

[40] Evaluation durch die studentischen Mitarbeitenden Emilie Buri und Tino Bruni.

[41] Vgl. dazu auch Anm. 46.

Geschlecht	Anzahl AutorInnen mit mind. 1 Text	Anteil am Gesamt in %
m	2'346	59.2
w	574	14.5
unklar	1'042	26.3
gesamt	3'962	100.0

Tabelle 3: Im Korpus vertretene Autorschaft[42]

Geschlecht	Anzahl Texte	
m	3'075	26.9
w	989	8.6
unklar/Anonymus	7'326	64.0
ohne Angaben	62	0.5
gesamt	11'452	100.0

Tabelle 4: Ins Korpus aufgenommene Texte nach Autorschaft

3.8 Regionale Verteilung

Als Referenzkorpus für die deutsche Standardsprache in Deutschland, Österreich, der Schweiz und Südtirol bildet das CHTK eine plurizentrische Sprache[43] ab, d. h. es enthält in seinen Texten und Sprachproben gleichermassen berechtigte und kodifizierte nationale und regionale Sprach-Varietäten. Entsprechend müssen variationslinguistische Fragestellungen an dieses Korpus herangetragen werden können. Das ursprüngliche Konzept, über die Lokalisierung der Autorschaft durch Angaben zu Geburtsort, Todesort und Lebensmittelpunkt-Ort eine möglichst präzise regionale Verteilung der Texte zu erreichen, hielt den Realitäten nicht stand, weil oft keine Aussagen darüber möglich waren. Methodisch hätte sich beispielsweise die Frage lösen lassen, welches denn für individualsprachliche Fragestellungen bzw. regionalsprachliche Einordnung der relevante 'Lebensmittelpunkt' eines Autors sein soll – die sprachliche Sozialisierung? die besuchten Schulen? der hauptsächliche Aufenthaltsort? Hätte man pragmatisch die individuelle Sprachsozialisierung und damit die Schulzeit fest als wichtigsten Einfluss festgelegt und grosszügig ignoriert, dass viele Autoren des 20. Jahrhunderts seit früher Kindheit mobil waren und somit mehrere Lebensmittelpunkte in Frage kämen, so wäre trotzdem die Tatsache bestehen geblieben, dass sich bei etwa rund der Hälfte der Texte im CHTK die Autorschaft nicht bestimmen lässt.[44] Dazu kam, dass die Partnerprojekte in Berlin und Wien, beide bei der Digitalisierung ihrer Texte weit fortgeschritten, dem Regionengedanken bisher wenig Aufmerksamkeit geschenkt hatten. Aufwendige Nachbearbeitung wäre also Pflicht geworden. Die regionale Verteilung wurde aus diesen Gründen auch beim CHTK ein 'weiches' Kriterium: Bei allen vier Werkkategorien, v.a. aber bei den journalistischen Texten und bei der Belletristik, wurde darauf geachtet, möglichst die gesamte deutsche Schweiz abzudecken und dies durch Angabe der Region in den Metadaten zu verankern.

[42] Stand: 28.2.2009.

[43] Vgl. Anm. 13.

[44] Von den in der Datenbank des STK enthaltenen 11'452 Einträgen zu Texten, die für die Aufnahme ins Korpus vorgesehen sind, haben 7'326 den Autoreneintrag 'Anonymus'; vgl. auch Tabelle 4.

Region	**Kantone**
CH-nordost	Schaffhausen, Thurgau (z.T.)
CH-nordwest	Aargau (z.T.), Basel-Landschaft, Basel-Stadt, Solothurn (z.T.)
CH-ost	Appenzell Ausserrhoden, Appenzell Innerrhoden, Zürich, Sankt Gallen, Thurgau (z.T.)
CH-süd	Graubünden, Uri, Wallis
CH-südost	Glarus, Graubünden
CH-südwest	Wallis, Bern (z.T.)
CH-west	Bern (z.T.), Freiburg (z.T.), Solothurn (z.T.), Aargau (z.T.)
CH-zentral	Luzern, Nidwalden, Obwalden, Schwyz, Uri, Zug

Tabelle 5: Einteilung der deutschsprachigen Schweiz in sprachliche Regionen[45]

Regionalspezifische Fragestellungen lassen sich derzeit im CHTK über den Publikationsort sowie über die Publikationsregion untersuchen, wobei natürlich die Suchresultate weiter analysiert und interpretiert werden müssen, insbesondere bei der Belletristik, wo vom Publikationsort selten auch auf die sprachliche Herkunft des Autors / der Autorin geschlossen werden kann. Bei der Journalistik und den Gebrauchstexten dagegen ist Publikationsort oder -region in dieser Hinsicht relativ aussagekräftig.

Die Suche im CHTK nach dem Wort *Anzug,* das in der Stadt Basel, ausser den gemeindeutschen Bedeutungen 'textiler Bezug', 'Kleid' und 'Beschleunigungsvermögen' die Sonderbedeutung 'Antrag im (Baselstädtischen) Parlament'[46] hat, ergab in der Region CH-nordwest 41 Treffer, wovon 14 Belege für die spezifisch Baslerische Bedeutung sind; in der Region CH-ost hat von 90 Belegen für *Anzug* kein einziger diese Bedeutung, in der Region CH-west keiner von 9 Treffern (die anderen Regionen haben das Wort nicht belegt).[47]

4 Möglichkeiten und Grenzen des CHTK

4.1 Allgemeines

Viele Arbeiten zur deutschen Standardsprache in der deutschsprachigen Schweiz mussten bisher ihre Daten fast von Grund auf selbst erarbeiten. Zwar berücksichtigen Institutionen wie das IdS[48] oder seit jüngerer Zeit das DWDS in ihren Korpora auch Texte von Schweizer AutorInnen, aber für die Zwecke der Schweizer Forschenden reichten die bereitgestellten Bestände an Texten mit Schweizer Herkunft nicht aus, sei es quantitativ oder qualitativ, weil viel Material ausschliesslich aus wenigen Zeitungen oder "nur" von international bekannten Schweizer AutorInnen stammte. Zudem liessen sich die Texte von Schweizer AutorInnen nicht isoliert betrachten.[49]

Das CHTK sollte hier Abhilfe verschaffen und stand dabei nicht unter dem Diktat, einer möglichst reichen Ausbeute für einen ganz bestimmten Forschungszweck dienen zu müssen. Als Korpus eines Zentralrepertoires konzipiert, waren es hauptsächlich zwei Ansprüche,

[45] Diese Einteilung wurde übernommen aus Ammon et al. (2004: XVIII); z.T. leicht modifiziert.

[46] Vgl. Ammon et al. (2004).

[47] Die Resultate beruhen noch auf dem Stand des STK vom Februar 2009, der ca. 90% der 20 Mio. Textwörter zur Verfügung stellt.

[48] www.ids-mannheim.de/.

[49] So hat etwa das *Variantenwörterbuch des Deutschen* (Ammon et al. 2004) eigene papierbasierte und elektronische schweizerische, österreichische und bundesdeutsche Korpora aufbauen müssen, mit deren Hilfe das Wörterbuch erarbeitet werden konnte. Die elektronischen Teile der Korpora umfassen rund 14 Mio. Wörter, die ganz im Hinblick auf die nationalen Varianten des Deutschen gesammelt worden sind. Der Aufbau des Schweizer Textkorpus unterlag gegenüber dem Variantenwörterbuch anderen Einschränkungen. Vgl. oben Kap. 3.

denen es zu genügen hatte: Spezifität in Bezug auf nationale (und sprachregionale) Varianten und Varietäten zum einen, historische Tiefe zum anderen.

Die Bedingungen, denen der Aufbau des Korpus unterworfen war, sind bereits weiter oben dargestellt worden. Im Folgenden sollen nun nach einem kurzen Vergleich der Ergiebigkeit des CHTK mit anderen deutschsprachigen Korpora anhand konkreter Abfragebeispiele die grundlegenden Möglichkeiten der Abfrage und der Nachbearbeitung der Resultate gezeigt und erläutert werden. Nicht eingegangen wird in diesem Kapitel auf die Möglichkeiten der Abfrage von Part-of-Speech-Tags.

Für das leichtere Verständnis der folgenden Abfragen werden hier vorab die Suchoperatoren, die in den folgenden Beispielen verwendet werden, aufgelistet. Die Einheit für Abfragen ist immer das Textdokument (in dem das Gesuchte durchaus mehrfach vorkommen und angezeigt werden kann).[50]

$l= Operator für Inanspruchnahme der Lemmatisierung bei der Abfrage

&& logisches UND (innerhalb des gleichen Textdokumentes)

|| logisches ODER (innerhalb des gleichen Textdokumentes)

! logisches NICHT (wird ignoriert, wenn es der einzige Operator ist)

* Platzhalter für keines, eines oder mehrere Zeichen, kann an beliebiger Stelle stehen

Distanzoperator, in Kombination mit einer Zahl (die für eine Anzahl Wörter steht), der Suchterm rechts des Operators muss innerhalb des angegebenen Bereichs rechts desjenigen Suchterms vorkommen, der vor dem Operator steht (z.B.: *melken #3 Euter)*

() zur Strukturierung komplexer Suchterme

"" für Phrasen (die auch den Lemma- oder Distanzoperator enthalten können); die einzelnen Suchterme müssen in der angegebenen Reihenfolge vorkommen.

4.2 Ergiebigkeit des CHTK im Vergleich mit Cosmas und DWDS

Bevor wir etwas ausführlicher anhand einiger konkreter Beispiele auf die Abfragemöglichkeiten des CHTK eingehen, soll hier zuerst etwas allgemeiner die Ergiebigkeit des CHTK im Vergleich mit anderen deutschsprachigen Korpora vorgestellt werden.

Einige Stichproben aus dem CHTK im Vergleich mit Cosmas-II[51] und DWDS-Kerncorpus müssten aufzeigen können, dass sich regionale Unterschiedlichkeiten in den einzelnen Korpora niederschlagen, wobei zu berücksichtigen ist, dass Cosmas-II mit dem St. Galler Tagblatt und dem Zürcher Tagesanzeiger viel (opportunistisches) Sprachmaterial aus der Schweiz beinhaltet.

Wenn wir zur Probe von einigen bekannten Helvetismen ausgehen, so finden wir Folgendes im Vergleich, wobei die Anzahl der Belege pro 100'000 Textwörter dargestellt wird (ausgehend von den auf den Homepages genannten Gesamtzahlen der Korpora, die durchsucht werden können).[52]

[50] Ausführliche Erläuterungen finden sich unter www.dwds.de/HilfeSuche (Stand 28.2.2009).

[51] Durchsucht wurde das Archiv *W-öffentlich – alle öffentlichen Korpora des Archivs W*, wobei das "Archiv" Texte "der geschriebenen Sprache" enthält.

[52] Die Abfrage des DWDS-Korpus und des Cosmas-II-Korpus bildet den Stand vom November 2008 ab. Die CHTK-Daten sind vom März 09 und beziehen sich auf die Gesamtgrösse von 20 Mio. Tokens.

	Cosmas-II	**DWDS**	**CHTK**
*parkiert**	(611 Belege) 0.06 pro 100'000	(4 Belege) 0.004 pro 100'000	(29 Belege) 0.15 pro 100'000
Spital oder *Spitäler*	(39'623) 3.58	(260) 0.26	(803) 4.02
Lemma *Krankenhaus*	(73'068) 6.61	(2'004) 2.0	(189) 0.95
*passier**	(113'299) 10.25	(3'203) 3.2	(831) 4.16
*verprofiantier**	(0) 0	(0) 0	(1) 0.01
*traversier**	(66) 0.01	(6) 0.01	(83) 0.42
*traktier**	(2'108) 0.19	(136) 0.14	(32) 0.16
*Traktand**	(5'971 / 34)[53] 0.54 / 0.003	(29) 0.03	(156) 0.78
Velo	(5'190 / 338)[54] 0.54 / 0.03	(6) 0.01	(120) 0.6
Lemma *ziemlich*	(59'067) 5.34	(6'612) 6.61	(1'941) 9.71

Tabelle 6: Stichproben zum Vergleich dreier deutschsprachiger Korpora[55]

Die sprachlichen Unterschiede – bezogen auf die Helvetismen – welche ausserhalb des Variantenwörterbuchs (Ammon et al. 2004) bislang oft nur aus dem Sprachgefühl heraus beurteilt wurden, spiegeln sich im exemplarisch durchgeführten Vergleich eindeutig statistisch wider. Eine erste Validierung der Brauchbarkeit des 20-Mio.-Korpus ist damit erbracht. Der Vergleich zeigt nebenbei aber auch, dass die unspezifische Zusammensetzung des Cosmas-II-Korpus beispielsweise bei variationslinguistischen Fragestellungen, wie angenommenen, zu grob ist. Die für *Traktand** und *Velo* durchgeführte Detailbetrachtung der rein statistischen Ergebnisse zeigt, dass die Aussage im grossen, über 1 Mrd. Tokens zählenden Korpus hinsichtlich der Fragestellung nahezu wertlos ist. Umgekehrt zeigt sich dadurch, dass zumindest bei spezifischeren Fragestellungen die kleineren Korpora aufgrund ihrer definierteren Zusammensetzung deutlich im Vorteil sind. Im Weiteren lassen wir die Ergebnisse aus Cosmas-II daher für diese Fragestellung beiseite.

Rein statistische Differenzen können schliesslich Anlass zu genaueren (bspw. semantischen) Betrachtungen der Ergebnisse bieten. So lässt sich am CHTK beispielsweise zeigen, dass sich *passieren* von der Bedeutung 'an etwas vorbeigehen' hin zu 'etwas geschieht' entwickelt, wobei die erste Bedeutung im Schweizerhochdeutschen wesentlich länger Bestand hat als in den Belegen im DWDS.

Im Vergleich mit dem DWDS zeigt sich auch in anderen Hinsichten, dass das CHTK zusätzliche Aspekte einbringt: Sucht man nach den Lemmata *Dialekt* und *Mundart*, erhält man im

[53] Lediglich 27 Belege (6 aus *Wikipedia*, 1 aus der *Zeit* und 27 aus österreichischen Tageszeitungen) stammen nicht aus Schweizer Tageszeitungen (*St. Galler Tagblatt* oder *Zürcher Tagesanzeiger*). Rechnet man diese heraus, ergibt sich der zweite angegebene Wert.

[54] Auch hier gestaltet sich die Situation ähnlich wie bei *Traktand**: 176 Belege stammen aus *Wikipedia* oder deutschen Tageszeitungen, 162 Belege aus österreichischen Tageszeitungen, der Rest aus Schweizer Tageszeitungen.

[55] Bei *Krankenhaus* und *ziemlich* handelt es sich unter den hier aufgeführten Beispielen nicht um Helvetismen .

CHTK 4,5-mal mehr Treffer pro 100'000 Tokens als im DWDS[56], wodurch sich – soziolinguistisch interessant – eine Präferenz des Themas in den Schweizerdeutschen Texten zeigen lässt. Ebenfalls auf den ersten Blick zeigt sich das mit 3.13 pro 100'000 Tokens deutliche Übergewicht von *Mundart* im CHTK (lediglich 1.41 pro 100'000 für *Dialekt*), während das Verhältnis im DWDS zwar nicht so deutlich aber genau umgekehrt ist.[57] Mit einer Detailbetrachtung der Verteilung in den einzelnen Werkkategorien lassen sich die signifikanten Unterschiede noch weiter spezifizieren: So entfallen über die Hälfte (56 Prozent) der *Dialekt*-Belege im CHTK auf die Kategorie Sachtexte, während im DWDS mit jeweils etwa 30 Prozent die Kategorien Gebrauchstexte und Belletristik den Hauptanteil ausmachen. Bei den *Mundart*-Belegen ist die Situation leicht verschieden. Im DWDS entfällt die Hälfte auf die Kategorie Sachtexte, während im CHTK die Kategorien Sachtexte (46 Prozent) und Gebrauchstexte (31 Prozent) über dreiviertel der Belege enthalten. Bezieht man auch noch die diachronen Informationen mit ein, fällt vor allem im DWDS auf, dass fast alle Belege[58] für *Mundart* aus Texten vor 1944 stammen. Es zeigt sich daran die unterschiedliche diachrone Entwicklung von *Dialekt* und *Mundart* im Deutschen und Schweizerhochdeutschen vor dem Hintergrund der Ereignisse des Nationalsozialismus. Es lohnt also durchaus, die statistischen Zahlen, die eine Korpusrecherche auswirft, beispielsweise bezüglich ihrer Verteilung auf die Jahrhundertviertel oder Werkkategorien zu betrachten. Die spezifischer konstituierten Korpora wie das CHTK sind bei solchen über die blosse Statistik hinausreichenden Aspekten eindeutig im Vorteil, was in erster Konsequenz auch das Spektrum an möglichen Fragestellungen erweitert. Wie sich am Beispiel *Mundart* vs. *Dialekt* zeigt, lassen sich mit dem CHTK beispielsweise soziolinguistisch-kulturelle, sprachgeschichtliche, semantische und viele weiter Untersuchungen anstellen. Der Vergleich mit dem DWDS und gegebenenfalls auch dem AAC bietet sich in allen Fällen nicht nur aufgrund der Parallelität der Teilkorpora des *Korpus C4* an.

Je kleiner ein Korpus ist, desto eher wird unterstellt, die Anzahl der Types reduziere sich mit der Menge der Tokens. Dass dem nicht so ist, wenn die Texte wie im CHTK nach qualitativen Kriterien ausgewählt wurden, zeigt ebenfalls der Vergleich mit dem fünfmal grösseren DWDS-Korpus. Die Suche nach seltenen Wörtern des Deutschen[59] wie *grein*, *weiland*, *anheischig*, *spornstreichs*, *Remise*, *Mär* macht deutlich, sie kommen entweder in keinem der beiden Korpora vor oder in nahezu identischer Häufigkeit. Das bedeutet einerseits methodologisch, dass das CHTK ausreichend gross dimensioniert ist, um auch seltene Wörter der Sprache abzubilden. Ausserdem bewährt sich nun, dass einem qualitativen Ansatz gegenüber einem quantitativen der Vorzug gegeben wurde. Andererseits bedeutet das Ergebnis inhaltlich, dass Schweizer Texte durchschnittlich nicht sprachbewahrender oder – wenn man so will – "altmodischer" sind als die Texte, die im DWDS repräsentiert sind.

Komplexe syntaktische Abfragen, welche mit dem CHTK möglich sind, haben wir hier aus Platzgründen ausgespart. Dennoch lassen sich auch mit Einzelwortabfragen morphologische und syntaktische Aspekte beleuchten. Beispielsweise lässt die signifikant häufigere Verwendung des Wortes *wo* im CHTK[60] auf eine deutliche Repräsentanz von syntaktischen Dialektphänomenen in den deutschschweizerischen Standardtexten im Korpus schließen. Auch die Artikel *der*, *die*, *das* sind statistisch häufiger im CHTK als im DWDS, was auf deren grössere Präsenz im Dialekt verglichen mit dem Gemeindeutschen zurückzuführen sein könnte (vgl. auch vor Eigennamen: *de Marcel – Marcel*).

---

[56] DWDS: absolut = 1051 Belege = 1.05 pro 100'000; CHTK: absolut = 907 Belege = 4.54 pro 100'000.

[57] *Mundart*: 0.4 pro 100'000; *Dialekt*: 1.05 pro 100'000.

[58] Lediglich für drei Belege aus Nachschlagewerken trifft das nicht zu.

[59] Wir haben diese von der Seite www.wortmuseum.de entnommen, wo nach eigenen Angaben vom Aussterben bedrohte Wörter "ausgestellt" werden.

[60] 78 Belege pro 100'000 zu 46 Belegen pro 100'000 im DWDS.

Nach diesem einleitenden Vergleich der Ergiebigkeit des CHTK mit anderen deutschsprachigen Korpora soll im Folgenden anhand einiger konkreter Beispiele auf die Abfragemöglichkeiten im CHTK eingegangen werden.

4.3 Historische Beispiele

Das erste Beispiel zeigt den Bedeutungswandel von *Ampel*: Während in allen Belegen bis 1957 die *Ampel* ein hängendes Licht meint (ursprünglich war damit das hängende ewige Licht in den Kirchen gemeint), so ist in allen Belegen ab 1971 die Verkehrsampel gemeint. Zeitgleich mit dieser neuen Bedeutung finden sich auch erste Belege für das Kompositum *Verkehrsampel.* Die Resultate wurden mit der einfachen Suchabfrage *Ampel* gewonnen. Die Abfrage nach allen Formen des Lemmas, *$l=Ampel,* hätte 31 statt nur 18 Belege ergeben. Durch das Korpussystem nicht von der Lemmatisierung erfasst wird jedoch die *Verkehrsampel*, sie muss separat gesucht werden, ist jedoch für das Verständnis der historischen Entwicklung wichtig. Mithilfe einer in das Web-Interface eingebauten Sortierfunktion können die Belege leicht auf- oder absteigend nach Jahr sortiert werden (in Abb. 1 aufsteigend; weitere Sortieroptionen sind *Autor* und *Satzlänge).* Die meisten Belege für *Ampel* stammen aus belletristischen und aus Gebrauchstexten, Sachtexte und Zeitungen sind kaum vertreten. Auffällig ist, dass die ältere Bedeutung von *Ampel* hauptsächlich in belletristischen Texten zu finden ist, während die neuere Bedeutung zusätzlich im Sachtext vorkommt. Dies lässt sich der Spalte ganz links entnehmen *(Bel* für Belletristik, *Geb* für Gebrauchstexte, *SaT* für Sachtexte, *Ztg* für journalistische Texte; der Inhalt dieser und der nebenstehenden Spalte lässt sich übrigens vor oder nach der Abfrage auch anders konfigurieren).

|<<	<	>	Treffer **1 - 18** von **18** || Abfrage: '**Ampel**' sortiert nach: **Jahr**, aufsteigend	<	>	>>|
SaT 1910			... mit Luft gefüllt hatten. Bei dem kärglichen Licht einer rußigen **Ampel** mußte zunächst die Wunde um einige Zentimeter erweitert, dann das ...			
Bel 1922			... hatte. Im warmen Flur, unter dem Licht der hochgezogenen **Ampel** , öffnete Doris ihren Pelz und ließ ihn lässig über die Schultern ...			
Bel 1922			... ihn seine Frau, die krank im Bette lag, als er noch spät bei der **Ampel** hockte und vor sich hin döste. Er fuhr auf: „ Gesehen habe ich ...			
Bel 1924			... die Sonne hing in der Glocke, wie eine große, leise schwankende **Ampel** . Der gepflasterte Platz vor der Kirche wurde von ihnen schwarz ...			
Bel 1950			... Arbeit übermannt hatte. Vor ihm flackerte in der frisch gefüllten **Ampel** nur noch ein schwaches Ölflämmchen. So mußte er einige Stunden ...			
Bel 1957			... und des Chores leuchteten ihm beim Scheine der ewigen **Ampel** die bunten Farben der Fresken entgegen, als blühten hier liebliche ...			
Bel 1971			... Minuten schon an der gleichen Stelle, vor einer Kreuzung ohne **Ampel** , der Chauffeur im Gespräch mit einem Mann, der mit einem ...			
Bel 1972			... Haas biegt in die Winterhaiderstraße ein. Wieder eine rote **Ampel** . Er hält an. ...			
Geb 1988			... Ampelsteuerung **Ampel** *Ampel* 1 2 Lauflicht ...			
Geb 1988			... ohne direkten Vergleich mit dem Istwert eigestellt(z.B. **Ampel** schaltet alle Minuten auf grün). Starkstrom ...			
Geb 1988			... jeder Verkehrsteilnehmer stellt, wenn er an eine eingeschaltete **Ampel** kommt:" Welche Farbe zeigt die *Ampel* an?" Es sind drei Antworten ...			
Geb 1988			... Frage kann auch in folgende Fragen zerlegt werden: Zeigt die **Ampel** grün an? ...			
Geb 1988			... Zeigt die **Ampel** gelb an? ...			
Geb 1988			... (Die dritte Frage," Zeigt die **Ampel** rot an", erübrigt sich) ...			
Geb 1990			... , den Hampelmann, dr Grättimaa, sowie das Männchen an der **Ampel** ? Wer wundert sich schon über Männchen machende Hunde(damen) oder ...			
Geb 1990			... die Norm, der Normmensch ist der Mann. Alle haben z.B. an der **Ampel** zu warten; das Männchen/der Mann steht für den Menschen schlechthin ...			
Geb 1990			... Vom Strichmännchen über den Schneemann bis zum Männchen an der **Ampel** u.a.m.: Sie alle haben in der gängigen" Muttersprache" keine ...			
Bel 1994			... da sah ich sie. Sie wartete brav vor der Kreuzung darauf, daß die **Ampel** auf grün schaltete. Sie saß auf dem Damenfahrrad, daß ich mir ...			

Abb. 1: Beispiel *Ampel*: Bedeutungswandel im CHTK

Ztg	1971	... auf einer Wolke stehend, schauen unzählige Kinder gebannt auf eine	**Verkehrsampel**	. Wann wird sie ihnen grünes Licht zur freien Fahrt auf die Erde ...
Ztg	1971	... gelungen, Adam zu betören. Damit schaltet oben auf der Wolke die	**Verkehrsampel**	auf Grün und sogleich schwärmen die Kinderlein aus, der Erde ...
Geb	1988	... Ampelsteuerung	**Verkehrsampeln**	gehören zum alltäglichen Strassenbild. Sie können durch feste ...
Ztg	1966	... Fuss, ein Ringlein, eine Haarsträhne tausendmal wichtiger als eine	**Verkehrsampel**	, als die Anrede eines Tramkondukteurs. Der Stil dieses Films ...
SaT	1978	... Gartenarchitektur durcheinanderbringen. Haben uns Fahrspuren und	**Verkehrsampeln**	so in den Bann geschlagen, dass wir das kulturelle Erbe übersehen? ...
Bel	1972	... die Fahrkarte nicht lösen, sagt sich Rau, nicht heute. Die grüne	**Verkehrsampel**	leuchtet auf und entfesselt Scharen von Fußgängern, die auf den ...
Bel	1972	... hatten sich an dieser Stelle einige Unfälle ereignet. Nun hat man	**Verkehrsampeln**	angebracht. Es hat aufgehört zu regnen. ...
Bel	1972	... Straße, sehr überschaubar in ihrer Bewegung. Die Kreuzung mit den	**Verkehrsampeln**	. Die Leute, verkürzt gesehen, merkwürdig untersetzt und ...
Bel	1972	... Er blickt unentwegt geradeaus. Eine	**Verkehrsampel**	zwingt ihn anzuhalten. Dann fahren sie den Schienen der ...
Geb	1988	... Codes und Codierung Die Frage:" Welche Farbe zeigt die	**Verkehrsampel**	an", lässt sich in zwei Elementarfragen zerlegen. Die Antwort ...

Abb. 2: Zusatz zu *Ampel*: *Verkehrsampel* im CHTK

Das zweite Beispiel zeigt eine neuere Entlehnung, deren Auftauchen sich im CHTK gut belegen lässt: *Disco* wurde laut dem *New Oxford American Dictionary* als Kurzwort in den USA gebildet. Ab 1982 taucht es im CHTK auf – im Gegensatz zum DWDS nur in der Schreibung *Disco* (im DWDS macht die Schreibung mit *k* einen Drittel der Belege aus). Abgefragt wurde das Lemma, als *$l=Disco*, deshalb tauchen zwei Pluralformen *Discos* auf. *Disco* kommt überwiegend in journalistischen Texten vor. Interessiert man sich für bestimmte Belegstellen näher, so lassen sich beliebige einzelnen Zeilen der KWIC-Darstellung per Mausklick "auf- und zuklappen". Dadurch werden mehr Kontext und eine genaue Quellenangabe sichtbar.

|<< < > Treffer **1 - 26** von **26** || Abfrage: **'$l=Disco'** sortiert nach: **Jahr**, aufsteigend < > >>|

Ztg 1982	... sie vollberuflich als Fotomodell tätig. Eher unauffällig in der	Disco	. Maria, in hautenge, schwarze Lederhosen und einen beigen ...

Als Überbrückung servierte sie, kam aber dann nicht mehr dazu, eine Lehre als Sehneiderin - dem weiteren Wunschberuf - zu ergreifen. Heute ist sie vollberuflich als Fotomodell tätig. Eher unauffällig in der Disco. Maria, in hautenge, schwarze Lederhosen und einen beigen Rollkragenpullover gekleidet.

pm (Aargauer Zeitung). 1982. «Für Maria dreht sich das Karussell». In: Aargauer Zeitung, 175.

Ztg 1984	... HOFFMANN KLAUS W., Wenn der Elefant in die	Disco	geht. Geschichten und Märchenlieder... ...
Ztg 1985	... so in die Stadt zu gehen, finde ich langweilig, und für die	Disco	fehlt mir das Geld. Am liebsten würde ich schon in einer glatten ...
Ztg 1985	... das die meisten zu simpel und veraltet. Heute geht man in die	Disco	, drückt auf Knöpfe oder glotzt stundenlang in den' Affenkasten'. ...
SaT 1987	... ist das Büro und hier finden immer wieder Veranstaltungen statt:	Discos	, Vollversammmlungen, Treffen mit anderen Gruppen. Die ...
SaT 1988	... man sich am Stammtisch, im Kirchenchor, im Sportklub oder in der	Disco	. Es ist mit anderen Worten die Umgebung, mit der wir verbunden ...

Hier kennen wir uns aus, hier können wir mitreden. Da trifft man sich am Stammtisch, im Kirchenchor, im Sportklub oder in der Disco. Es ist mit anderen Worten die Umgebung, mit der wir verbunden sind.

Domeyer, Barbara (et al.). 1988. «Wo Werbung am nächsten kommt». (S. 20)

Ztg 1989	... mitgestaltet, bis hin zum unverbindlichen Treffpunkt oder	Disco	. Beides, ganz offene und ganz verbindliche Jugendarbeit, kann in ...
Ztg 1989	... immer weniger, wie mir die Leiter des Zentrums bestätigen. Mit	Disco	und offenen Bädern, mit Videofilmen und gemeinschaftlichem ...
Ztg 1989	... Im GZ Riesbach herrscht ebenfalls Flaute. Hier gibt es keine	Disco	, kein Unterhaltungsangebot, « nur » heisse Schokolade und ...
Ztg 1989	... aus Lautsprechern, Weihnachtsterne baumeln von der Decke: eine	Disco	eben, nicht mehr, nicht weniger. Beim Hinausdrängen komme ich an ...
Ztg 1989	... der Limmat zu suchen: Das « Mascotte » lockt mit frühmorgendlicher	Disco	. Vor dem Eingang drängen sich Dutzende von Mädchen und Burschen, ...
Ztg 1992	... Fr, Sa 20-0.30 Uhr Frauenfest in der Kammgarn mit Film,	Disco	, Konzert; 4.9., 20 Uhr St. Gallen ...
Ztg 1992	... Bern	Disco	jeden letzten Sa im Monat ab 20 h, Frauenzentrum, Langmauerweg 1 ...
Ztg 1992	... Lesbenbar: Do, ab 20 h	Disco	« Lesben laden zum Tanz » jeden 1. Sa im Monat ab 22 Uhr ...
Geb 1992	... Sonst bin ich immer allein. Ich will nicht in die	Discos	, weil mich sowieso niemand beachtet. Aber manchmal - selten - ...
Geb 1992	... völlig deprimiert. In diesem Zustand ging ich dann noch in die	Disco	. Aber ich hielt es einfach nicht aus vor dem Fernseher im ...
Geb 1992	... Ich beschäftigte mich lieber mir klassischer Musik als mit	Disco	, besuchte eher ein Museum als eine Bar, las lieber ein Buch und ...
Ztg 1993	... steht nichts in Ihrer Bestellung! »). Und nervt sich über die	Disco	, die bis sechs Uhr morgens neben dem Schlafzimmer ihre ...
Ztg 1995	... Kreisen 4 und 5 illegal gewirtet wird. Viele illegale Bars und	Discos	, welche notabene dem Staat keine Abgaben bezahlen, geschweige ...
Ztg 1997	... sieht die Zukunft der Welle aus? Obwohl ich als Gründerin dieser	Disco	sehr an dieser Arbeit hänge, bin ich aus beruflichen Gründen ...
Ztg 1997	... Rahmen die Weiterbestehung zu wahren und die bis dahin kleinere	Disco	zu erweitern und aufzubauen. Ab September 1997 wird die neue ...
Ztg 1997	... Arme gegriffen haben und mich unterstützt haben, damit ich diese	Disco	aufbauen und weiterbetreiben konnte. Auch weiterhin hoffe ich, ...
Ztg 1997	... Gaskessel Bern und von Christine Jost war aus der Idee bald eine	Disco	entstanden. Zu Beginn belegten wir einen kleineren Atelier-Raum ...
Ztg 1997	... kulturelle Arbeit zu finden. Durch die offene Haltung dieser	Disco	, die jedes Interesse versucht aufzunehmen, wird gegen starre ...
Ztg 1997	... SA 13.9. Oldies	Disco	Frauenraum Reithalle, Bern ...
Ztg 1998	... werden bei praktischen Übungen schwierige Lichtsituationen(von	Discos	bis zu Kirchenschiffen und Kanalisationen) gemeistert. Der Abend ...

Abb. 3: Beispiel *Disco* 'Tanzlokal, Tanzveranstaltung' (Entlehnung aus dem Englischen) im CHTK

In ihrer Lizentiatsarbeit hat Emilie Buri, studentische Mitarbeiterin des CHTK, im Kontext einer kulturlinguistischen Untersuchung eine Reihe von Verben im Hinblick auf ihre semantisch-syntaktische Einbettung und Verwendung im Laufe des 20. Jahrhunderts untersucht, Kollokationsanalysen vorgenommen und die Resultate mit analogen Analysen im DWDS verglichen (Buri 2008). Augenblicklich sind im CHTK noch keine Tools für Kollokationsanalysen implementiert. Die Texte des Korpus lassen sich aber lokal mit anderen Tools analysieren (in diesem Fall: *antconc*, www.antlab.sci.waseda.ac.jp). Das besondere Augenmerk war darauf gerichtet, welcher Sachgruppe ein Beleg angehört und ob bzw. wie sich diese Zugehörigkeit im Laufe der Zeit weiterentwickelt. Ein von Buri untersuchtes Verb ist *arbeiten.* Sie stellt dazu in einer ersten Zusammenfassung ihrer Analysen fest:

> Somit kann vorerst von einem (wohl eher sanften) Bedeutungswandel von *arbeiten* vor allem in den Sachgruppen Soziologie, Gesellschaft, Arbeit, Sozialgeschichte' Religion, Sport' Geschichte' Wirtschaft, Verkehr, Umweltschutz, Raumordnung' Schrift, Buch, Presse, Musik und Kultur, Erziehung, Bildung, Wissenschaft' ausgegangen werden. Interessant wäre es nun, die Geschichte der jeweiligen Sachgruppen zu untersuchen, um festzustellen, ob auch aussersprachlich ein Wandel stattgefunden hat. (Buri 2008: 45)

Und weiter:

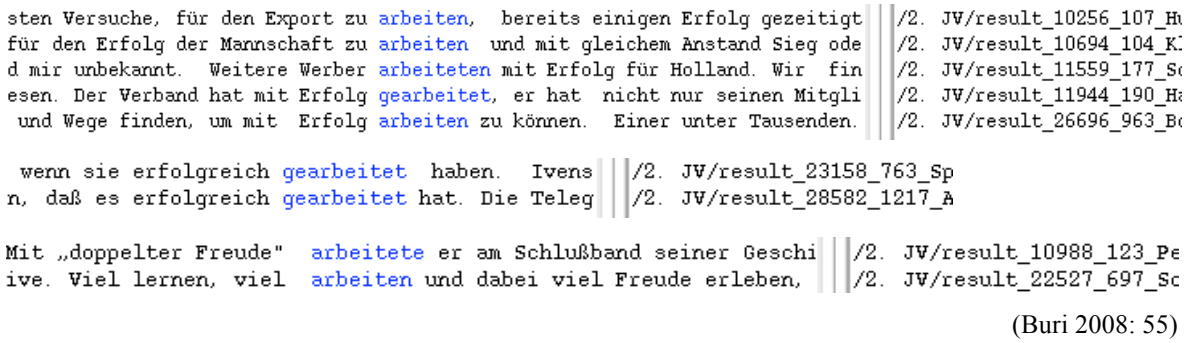
> Die Kollokate (als *tokens)* zum Verb *arbeiten* verändern sich in den Sachtexten des Schweizer Hochdeutschen zwischen den vier Jahrhundertvierteln des 20. Jahrhunderts und beeinflussen das Verb in seinem Signifikat kulturbedingt und kulturbedingend. (Buri 2008: 50)

Schliesslich:

> Für die Zeitspanne von 1925–1949 lassen sich […] einige spezielle Kollokate festmachen, die in den übrigen Jahrhundertvierteln nicht auftauchen. So tritt das Lexem *arbeiten* oft mit Berufswörtern auf. Diese bezeichnen beinahe ausnahmslos traditionelle, handwerklich orientierte Berufe. (Buri 2008: 54)

Ausserdem kookkurriert das Lexem *arbeiten* im zweiten Jahrhundertviertel mit Wörtern, die als positiv einzustufen sind. Diese Kookkurrenzen fallen auf, weil die Kollokate durch das ganze Jahrhundert hindurch grösstenteils neutrale, wenn nicht negative semantische Konnotationen aufweisen. Sie haben auch meistens mit dem (aufwändigen bis mühsamen) Arbeitsprozess, nicht aber mit dem erfolgreichen Endresultat zu tun (vgl. Buri 2008: 67f.).

> Die Bedeutung dieser für das zweite Jahrhundertviertel spezifischen Kollokate lässt sich als 'gewinnbringend und energiestiftend' zusammenfassen:
>
> 1925–1949: *Erfolg*, *erfolgreich*, *Freude*



(Buri 2008: 55)

Hier zeigt sich, dass trotz des (noch) geringen Umfanges des CHTK sich mit häufigen Wörtern (wie den von Buri untersuchten Verben) komplexe historisch ausgerichtete Analysen durchführen lassen, die Ergebnisse bis auf die Ebene der Sachgruppen zeitigen.

4.4 Spezifität hinsichtlich nationaler Varianten

In einem weiteren Block von Beispielen soll veranschaulicht werden, inwiefern das CHTK den Anspruch nach Spezifität in nationaler und regionaler Hinsicht einlösen kann, nachdem es ausschliesslich aus Texten besteht, die in der Schweiz produziert worden sind. Die historische und die national-regionale Komponente sind oft verschränkt, wie in den folgenden Beispielen deutlich wird.

Am Beispiel von *parkieren* und *parken* lässt sich schön zeigen, wie sich im Laufe des 20. Jahrhunderts diese Lehnwörter aus dem englischen *to park* im Deutschen etabliert haben. Allerdings nicht überall in der gleichen Art und Weise.

Anhand des CHTK lassen sich die quantitativen Anteile der beiden Varianten eruieren. Für Formen von *parkieren* finden sich 44 Belege, für solche von *parken* neun, also ein Verhältnis

ca. 5:1 *(parkieren : parken)*; im DWDS ist das Verhältnis 75:1 *(parken : parkieren*; bei einer Belegzahl von über 300). Alle Belege im CHTK stammen aus der Nachkriegszeit, während im DWDS die Belege schon 1925 einsetzen.

Die Lemma-Abfrage mit *$l=parkieren* ergab im CHTK kein befriedigendes Resultat, da der Lemmatisierer *parkieren* offensichtlich (noch) nicht kennt. Es wurden lediglich sieben Belege gefunden, alle mit dem Infinitiv. In diesem Falle konnte auf die Suche mit Platzhalter zurückgegriffen werden: *parkier** ergab die genannten 44 Fundstellen mit verschiedenen Flexionsformen.

Bel	1947	... Auto, Auto, dachte er fiebrig. Drüben bei der Kirche	parkierten	welche. Einem hakenschlagenden Hasen gleich flog er über den ...
SaT	1952	... Bäumen, den ungeheuren Schilfbüscheln und den merkwürdigen Büschen	parkiert	, dann das Zelt aufgestellt und ein Kochplatz eingerichtet. Jeden ...
Ztg	1955	... jetzt ein Stück zurück, um den Wagen für die Affen unsichtbar zu	parkieren	. Das schloss jedoch nicht aus, dass ein Späher der Herde schon ...
Geb	1960	... 50 Halbstarke zusammen und begannen mit Stöcken und Steinen auf	parkierte	Autos einzuhämmern. Die Polizei musste eingreifen und nahm ...
Geb	1960	... ihren Wogen tragend. ... und begannen mit Stöcken und Steinen auf	parkierte	Autos einzuhämmern. Man steigert sich in den Tumult hinein, weil ...
Bel	1964	... Tempo, drückte sanft auf den Gashebel und schoß um die Ecke. Er	parkierte	den Wagen und telephonierte seiner Frau; sie verabredeten sich im ...
Geb	1967	... von Oakland, der in San Francisco Einkäufe besorgt, falsch	parkiert	, so findet er wie bei uns unter dem Scheibenwischer einen ...
SaT	1970	... Straßen nach Dulliken fuhr, wo wir den Wagen vor dem Bahnübergang	parkierten	. Leise schlichen wir einer Baumallee entlang auf die Korkfabrik ...
Bel	1971	... schon in der Tür. Nur nicht so eilig, zuerst müssen die Wagen	parkiert	werden, die Leute aussteigen, nur nicht so eilig, viele brauchen ...
Bel	1972	... hat recht », sagt Lili, als sie an den funkelnden Karosserien der	parkierten	Autos vorübergehen. « Es geht wirklich nicht mehr ohne Wagen. ...
Bel	1972	... Wochen also. Langsam geht er dem Platz zu, auf dem er sein Auto	parkiert	hat. Ines steht auf einem Küchenschemel. ...
Geb	1973	... zu leiden hätten Dass die Arbeiter in den Aussenquartieren	parkiert	sind Dass das klassische Schemas Propaganda - Terrorismus - ...
Bel	1975	... der erste Käufer kommt. Ein Karren rattert durch die Gasse und	parkiert	neben einem blauen Fiat, während der Metzger eifrig Koteletts ...
Bel	1975	... Ausschau. Ihr Blick dringt durch Berti, Blumenvasen, Hausmauer,	parkierte	Autos, dringt durch Räume, Tausende von Kilometern weit. « Ihr ...
Bel	1977	... Ich blickte sie verständnislos an. « Wegen der Wagen, die da	parkieren	. » « Ja so. » ...
SaT	1978	... schieres Ärgernis am Ende eines stachligen Geländers auf dem durch	parkierende	Autos total entwerteten Platz. Wir schlugen einmal vor, die ...
Bel	1979	... kann? Für Fr. 20.- darf man einmal außerhalb markierter Felder	parkieren	, für Fr. 30.- sogar auf dem Seitenstreifen einer Autobahn anhalten ...
Bel	1979	... umziehen und bei der Haustür bereithalten - man kann hier nirgends	parkieren	. Ich gab Yuriko meinen zweiten Wohnungsschlüssel und ließ sie ...
Bel	1979	... sich der Randstein der gegenüberliegenden Straßenseite mit dem	parkierten	Auto, Renault 16, dunkelrot, es war zum Verrücktwerden. Die Leute ...
Bel	1985	... Luft, Landluft, tiefes Einatmen. Eiligen Schrittes zum nahe	parkierten	Auto, wie auf der Flucht. Verabschieden von der Aufseherin, mit ...

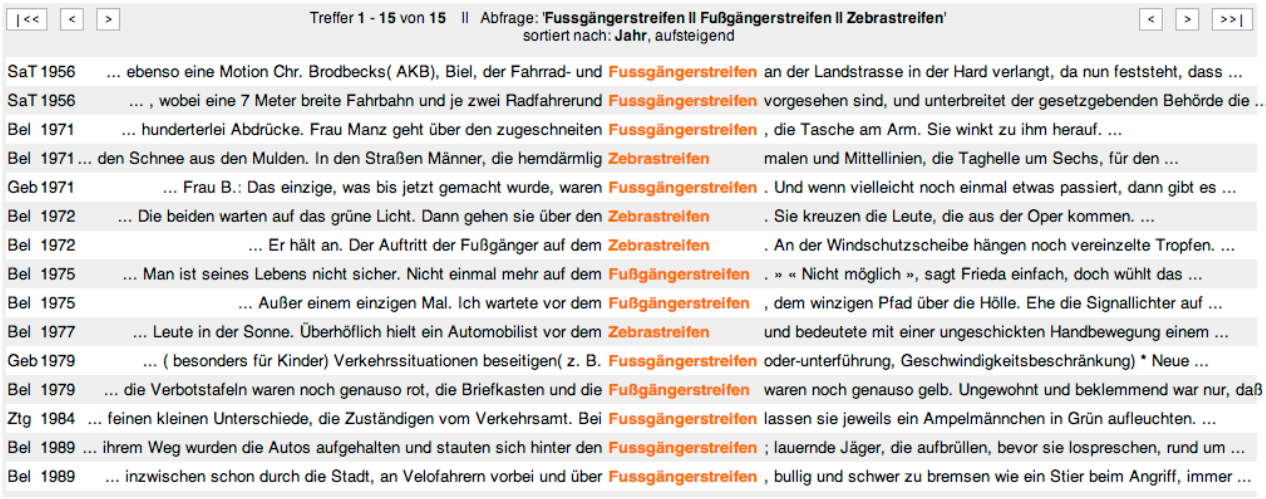
Abb. 5: Beispiel *parkieren* (CHTK) (44 Fundstellen, 20 dargestellt)

Demgegenüber erscheint das Verb *parken* weit seltener im CHTK (hier funktioniert die Lemmatisierung (Abfrage *$l=parken)*, jedoch mit einem falschen *Park* in den Resultaten, den man durch zusätzliche *und* und *nicht*-Operatoren ausschliessen kann: *$l=parken && !Park)*. Ein Beleg aus den insgesamt 9 ist metasprachlich und thematisiert u.a. genau die Differenz zwischen *parkieren* und *parken* (Abb. 6, letzte Zeile; der Beleg ist logischerweise auch in den 33 Fundstellen für *parkieren* enthalten).

Ztg	1952	... Ausflügler das Essen zubereiten können. An Sonntagnachmittagen	parken	vor diesen Plätzen Wagen an Wagen, und Tausende von naturliebenden ...
SaT	1972	... und bestehen auf Symmetrie. Winzig übersprungen, schief, vom Beruf	geparkt	, nistet die Vorfahrt. Ihr ist es Frost, Mühsal, das Schöne zu ...
SaT	1972	... fassen Beschlüsse und bestehen auf Vorfahrt. Regelwidrig	geparkt	, winzig, vom Frost übersprungen, nistet die Armut. Ihr ist es ...
Bel	1977	... heraus. « Haben Sie gesehen, wie die da vor unserem Gitter	parken	? » Tatsächlich hatte ich darauf gar nicht geachtet. ...
Bel	1988	... Zächliwil lag vor mir, als das Tal sich weitete. Ich	parkte	den Wagen am Bahnhof und ging direkt in die Post, wo ich mich nach ...
Bel	1988	... bis Zächliwil, wo ein ruhiger Regen fiel. Vor dem Gemeindehaus	parkte	ich den Wagen auf einem reservierten Feld, weil alle anderen ...
Bel	1994	... der Kopf von Jasmin Weber sichtbar wurde, lächelte ich milde. Sie	parkte	ein paar Meter oberhalb des Cafés. Sie wand sich aus dem dunklen ...
Bel	1994	... Hugentobler kam von der anderen Seite. Er	parkte	seinen Wagen auf einem kleinen Kehrplatz. Er war allein. ...
SaT	1997	... zeigen als das Binnendeutsche: grillieren, parkieren für grillen,	parken	oder Unterbruch, Zusammenzug, Wissenschafter für Unterbrechung, ...

Abb. 6: Beispiel *parken* (CHTK) (9 Fundstellen)

Ein ähnliches Bild zeigt sich bei den Wörtern *Fussgängerstreifen* und *Zebrastreifen*. Ihre Verteilung in den verschiedenen nationalen Kontexten ist ungleich. Für die Abfrage wurde der *Oder*-Operator || verwendet. Im CHTK finden sich 15 Belege, wovon 12 *Fuss/ßgänger-streifen* und 3 *Zebrastreifen*, was ein Verhältnis von 4:1 ergibt.



	<< < >	Treffer **1 - 15** von **15** || Abfrage: '**Fussgängerstreifen ll Fußgängerstreifen ll Zebrastreifen**' sortiert nach: **Jahr**, aufsteigend			< > >>|
SaT	1956	... ebenso eine Motion Chr. Brodbecks(AKB), Biel, der Fahrrad- und	Fussgängerstreifen	an der Landstrasse in der Hard verlangt, da nun feststeht, dass ...	
SaT	1956	... , wobei eine 7 Meter breite Fahrbahn und je zwei Radfahrerund	Fussgängerstreifen	vorgesehen sind, und unterbreitet der gesetzgebenden Behörde die ...	
Bel	1971	... hunderterlei Abdrücke. Frau Manz geht über den zugeschneiten	Fussgängerstreifen	, die Tasche am Arm. Sie winkt zu ihm herauf. ...	
Bel	1971	... den Schnee aus den Mulden. In den Straßen Männer, die hemdärmlig	Zebrastreifen	malen und Mittellinien, die Taghelle um Sechs, für den ...	
Geb	1971	... Frau B.: Das einzige, was bis jetzt gemacht wurde, waren	Fussgängerstreifen	. Und wenn vielleicht noch einmal etwas passiert, dann gibt es ...	
Bel	1972	... Die beiden warten auf das grüne Licht. Dann gehen sie über den	Zebrastreifen	. Sie kreuzen die Leute, die aus der Oper kommen. ...	
Bel	1972	... Er hält an. Der Auftritt der Fußgänger auf dem	Zebrastreifen	. An der Windschutzscheibe hängen noch vereinzelte Tropfen. ...	
Bel	1975	... Man ist seines Lebens nicht sicher. Nicht einmal mehr auf dem	Fußgängerstreifen	. » « Nicht möglich », sagt Frieda einfach, doch wühlt das ...	
Bel	1975	... Außer einem einzigen Mal. Ich wartete vor dem	Fußgängerstreifen	, dem winzigen Pfad über die Hölle. Ehe die Signallichter auf ...	
Bel	1977	... Leute in der Sonne. Überhöflich hielt ein Automobilist vor dem	Zebrastreifen	und bedeutete mit einer ungeschickten Handbewegung einem ...	
Geb	1979	... (besonders für Kinder) Verkehrssituationen beseitigen(z. B.	Fussgängerstreifen	oder-unterführung, Geschwindigkeitsbeschränkung) * Neue ...	
Bel	1979	... die Verbotstafeln waren noch genauso rot, die Briefkasten und die	Fußgängerstreifen	waren noch genauso gelb. Ungewohnt und beklemmend war nur, daß	
Ztg	1984	... feinen kleinen Unterschiede, die Zuständigen vom Verkehrsamt. Bei	Fussgängerstreifen	lassen sie jeweils ein Ampelmännchen in Grün aufleuchten. ...	
Bel	1989	... ihrem Weg wurden die Autos aufgehalten und stauten sich hinter den	Fussgängerstreifen	; lauernde Jäger, die aufbrüllen, bevor sie lospreschen, rund um ...	
Bel	1989	... inzwischen schon durch die Stadt, an Velofahrern vorbei und über	Fussgängerstreifen	, bullig und schwer zu bremsen wie ein Stier beim Angriff, immer ...	

Abb. 7: Beispiele Fuss/ßgängerstreifen und Zebrastreifen im CHTK

Demgegenüber finden sich im DWDS-Korpus nur 2 Belege für *Fußgängerstreifen*, jedoch 27 für *Zebrastreifen*. Vier davon (1, 2, 28, 29) haben eine andere Bedeutung als 'durch breite, weisse Streifen auf einer Fahrbahn markierte Stelle, an der die Fussgänger beim Überqueren Vortritt haben', und zwar die historisch frühesten beiden, was, zusammen mit den Belegen aus dem CHTK, darauf hindeutet, dass über die Sache im oben genannten Sinn vermutlich erst in den 1950er-Jahren geschrieben wurde.

Der regional-nationale Befund ist hier noch klarer als bei *parkieren/parken: Fussgängerstreifen* ist im CHTK eindeutig die Hauptvariante (4:1), während es im DWDS umgekehrt ist: hier ist *Zebrastreifen* die Hauptvariante (ca. 13:1).

1	Ge	1940	... Hast du jemals bei den Reklameröhren solche	Zebrastreifen	gesehen? -" Nein - denn sie werden mit ...
2	Ge	1943	... Brief von Arne Egebring mit den braunen und blauen	Zebrastreifen	der Zensur. Ärztliche Ratschläge. ...
3	Ze	1956	... in angemessener Weise zu ermöglichen. Vor	Fußgängerstreifen	hat der Fahrzeugführer langsam zu fahren und ...
4	Ze	1956	... anzuhalten. Auf besonders gekennzeichneten	Fußgängerstreifen	hat der Fußgänger den Vortritt. An den ...
5	Ge	1956	... nur auf den Fußgängerüberwegen, den sogenannten "	Zebrastreifen	". - Auch hier ginge alles glatter und ohne ...
6	Ge	1956	... auch ihr "Kavaliere an der Lenkstange". - Die "	Zebrastreifen	", jene Fußübergänge, auf denen der Fußgänger ...
7	Ge	1956	... Winke. Keinesfalls aber brausen wir an diese	Zebrastreifen	heran, um dann unmittelbar davor mit ...
8	Ge	1956	... bieten handsignierte Autorenexemplare. ·	Zebrastreifen	- · An einen Fußgängerüberweg fahren wir ganz ...
9	Ge	1956	... Winke. Keinesfalls aber brausen wir an diese	Zebrastreifen	heran, um dann mit quietschender Bremse ...
10	Ge	1957	... Das gilt auch für Fußgänger-Überwege (	Zebrastreifen	). - Wenn die Fahrgäste einer Straßenbahn zum Ein- ...
11	Ge	1957	... auf einem Fußgänger-Überweg (Schild oder/ und	Zebrastreifen	) erkennbar überschreiten wollen, das Überqueren ...
12	Ge	1965	... dort Personen aussteigen. Fußgänger, - die auf	Zebrastreifen	- die Fahrbahn überqueren, genießen erhöhten ...
13	Ge	1965	... darf deshalb nur mit mäßiger Geschwindigkeit an	Zebrastreifen	heranfahren, muß Fußgängern das Überqueren ...
14	Ge	1965	... - c) auf und näher als 5 m vor Fußgängerüberwegen (	Zebrastreifen	). - Ist man infolge eines Fahrzeugdefekts zum ...
15	Ge	1965	... Zum Schutze des Fußgängers sind Fußgängerüberwege (	Zebrastreifen	) auf der Straße markiert. Außerdem kann das ...
16	Ge	1965	... Außerdem kann das Hinweiszeichen auf den	Zebrastreifen	aufmerksam machen. Fußgänger sind ...
17	Ge	1965	... Fußgänger sind verpflichtet, nach Möglichkeit auf	Zebrastreifen	die Fahrbahn zu überqueren. Deshalb genießen ...
18	Ge	1965	... den Fußgängern das Überqueren der Fahrbahn auf	Zebrastreifen	ermöglichen. Sie dürfen deshalb nur mit ...
19	Ge	1965	... (dadurch natürlich auch Parkverbot). Hinter den	Zebrastreifen	darf man jedoch sofort wieder halten und parken. ...
20	Ge	1965	... anhalten. Das gilt auch dann, wenn keine	Zebrastreifen	den Vorrang des Fußgängers an Kreuzungen und ...
21	Ge	1965	... Schutz des Fußgängers genügt das Vorhandensein des	Zebrastreifens	, mit oder ohne das blaue Hinweiszeichen. 4. ...
22	Ge	1966	... nicht behindert werden. An Fußgängerüberwegen (	Zebrastreifen	im Volksmund) hat der Fußgänger das Vorrecht. ...
23	Ge	1966	... die vor Überwegen halten, wenn Sie den ganzen	Zebrastreifen	übersehen können. Überholen Sie vor einem ...
24	Ge	1966	... der Autobahn und weniger als 5 Meter vor	Zebrastreifen	. U. a. ...
25	Ge	1967	... paßt genau auf einen runden Kräcker -. -	Zebrastreifen	: Pumpernickelscheiben mit Butter bestreichen, mit ...
26	Be	1971	... ihre Füße auf einen	Zebrastreifen	setzen, und wenn wir ...
27	Ge	1986	... Überqueren von Straßen zum Erlebnis - selbst auf	Zebrastreifen	, die man doch zügig überqueren sollte. Wie ...
28	Wi	1989	... goldene Maske (1968), die Mini-Oper Es gibt doch	Zebrastreifen	für 2 Sänger und 8 Instr. 1970), die ...
29	Ze	1999	... Menschen darin kommen zum Vorschein. Shorts mit	Zebrastreifen	, Shorts mit Karos und mit Schottenmuster. ...

Abb. 9: Beispiel *Fußgängerstreifen* und *Zebrastreifen* im DWDS

4.5 Komplexere Abfragen und Filter

Die bisher dargestellten Beispiele kamen mit verhältnismässig einfachen Abfragen aus. Es sind jedoch auch komplexere Abfragen möglich, die beispielsweise das Auffinden von bereits bekannten oder vermuteten Kollokationen oder Bigrammen ermöglichen. In Abb. 10 ist ein Beispiel für die Eruierung von *staubig* im Zusammenhang mit *Weg* und *Strasse* dargestellt. Gesucht wird nach zwei Termen: Formen von *staubig*, bei denen mit maximal 3 Wörtern Abstand rechts eine Form von *Weg* steht, oder dann Formen von *staubig*, bei denen auch mit maximal drei Wörtern Abstand rechts eine Form von *Strasse* steht (Suchoperator *#3*) Diese Abfrage ergibt 17 Fundstellen.

<<	<	>	Treffer **1 - 17** von **17** || Abfrage: **"$l=staubig #3 $l=Weg" \|\| "$l=staubig #3 $l=Straße"** ' sortiert nach: **Jahr**, aufsteigend			<	>	>>|
SaT	1906		... Keine Schlangen, überall die üppigste Vegetation und selten	staubige	*Wege*; denn es vergeht selten ein Tag, an dem es nicht ein wenig ...			
Geb	1907		... Auseinanderbröckeln des Zuges sind von Anfang an streng zu rügen.	Staubigen	*Straßen* gehe aus dem Wege und lasse zur Vermeidung von Staub die ...			
SaT	1909		... noch Domodossolla. Immer den Felswänden nach schlingt sich die	staubige	und schlechte *Straße* in gewaltigen Krümmungen, die oft beinahe ...			
Geb	1920		... wurde. Beim Meiden aller Schäden(Übertreibungen, Befahren	staubiger	*Straßen*, zu starkes Vorwärtsneigen) kann es auch Frauen zur ...			
Bel	1924		... später töten wird. Mit mir selbst und dem Leben im Streit bin ich	staubige	*Wege* gereist, ohne abendliches Ziel, ohne Mutter, Vaterhaus und ...			
Bel	1925		... Mutter noch bis zur Fabrik hinunter, die weiß und grell an der	staubigen	*Straße* wartete. Männer und Frauen, auch solche, die noch wie ...			
Bel	1937		... den Anschein einer fraulich umsorgenden guten Seele. „ Ihr seid	staubig	von der *Straße*, Herr. Wie wär's mit einem Bad, das Euch erlabte?" ...			
Bel	1942		... gehörte ja auf einen Baum oderStrauch. Wie ist das wohl auf diese	staubige	*Strasse* gekommen? Hoffentlich ist es nicht verwundet!" ...			
Bel	1943		... und zwitscherten und die Spatzen durchsuchten die Roßbollen in	staubigen	*Straßen*. Mit offenem Mund starrte ich staunend in die herrliche ...			
Geb	1947		... So verursacht derselbe bei anstrengender Bewegung der Pferde auf	staubigen	*Straßen* Entzündungen der Augenschleimhäute, sowie ...			
Geb	1947		... sei aber auch hervorgehoben, daß in allzu kurzem Tempo auf harten	staubigen	*Straßen* ausgeführte Märsche für Pferd und Reiter ermüdend, ...			
Bel	1953		... zwingen. Nach einer schlaflosen Nacht wandert Matz auf	staubiger	, brennheißer *Straße* dahin; er wollte sich mit der Landschaft ...			
Ztg	1955		... Generation » des Französischen bediente. Über die ausgefahrene	staubige	*Strasse* erreichten wir bald die Stadt, wo ich im Gebäude des ...			
Bel	1955		... Bub hat einen Onkel Gerold. Was jetzt wohl der Käfer von der	staubigen	*Straße* macht? Maikäfer müßten ja gefangen und verbrannt werden, ...			
Ztg	1958		... kranken Menschen zu Tode geschunden hatte. Irgendwo auf den	staubigen	*Strassen* des christlichen Abendlandes sind an diesem Tage zwei ...			
SaT	1985		... auch der übliche schwarze Rock sind für längere Märsche bald auf	staubiger	oder kotiger *Straße*, bald durch nasse Wiesen und Sümpfe, über ...			
Bel	1991		... vergeblich, sich auf die Vorderläufe zu erheben und von der	staubigen	*Strasse* weg in die nahe Wiese zu kriechen. Sie wollte ihm helfen, ...			

Abb. 10: Komplexere Suchabfrage *staubige Strasse / staubiger Weg*

Ein in den oben stehenden Beispielen nicht angewendetes Tool sind die Filtermöglichkeiten, die auf dem Web-Interface des CHTK zur Verfügung stehen. Sie bieten die Möglichkeit, ein Set von Filtern auf eine Ergebnismenge anzuwenden. So kann die zeitliche Erstreckung eingeschränkt werden, es kann nach bestimmten AutorInnen oder Titeln gefiltert werden, die Ergebnismenge kann auf Texten aus bestimmten Sachgruppen, Werkkategorien, Produktionsregionen oder Produktionsorten eingeengt werden. Viele dieser Filtermöglichkeiten sind jedoch erst sinnvoll anzuwenden, wenn die Anzahl von zuvor gefundenen Belegstellen hoch ist oder wenn man sich beispielsweise für ein ganz bestimmtes Werk aus dem Korpus interessiert.

Ein Beispiel für die Anwendung des Filters soll hier die Abfrage *$l=Heu || $l=melken*. Die Abfrage dieser beiden Wörter ergibt insgesamt 338 Treffer. Mit dem Filter-Tool kann man sich nun Belege beispielsweise aus einzelnen Sachgruppen geben lassen. So finden sich knapp 9% der Belege in der Sachgruppe *32: Landwirtschaft, Garten* (bei gleichmässiger Verteilung dürften es lediglich knapp 3% sein, wie etwa in Sachgruppe 35: Spiel, Unterhaltung). Demgegenüber finden sich etwa in den Sachgruppen *3: Religion, 4. Philosophie, 31: Technik*, *21: Astronomie, Weltraumforschung* oder *20: Physik* (und in weiteren Sachgruppen) überhaupt keine Belege. Die verschiedenen Filtertools auf der Weboberfläche des CHTK sind in ihrer Reichhaltigkeit und ihrer Ausrichtung auf Metadaten der Korpustexte Ausdruck der konsequenten Ausrichtung auf eine grosse Vielfalt der Texte.

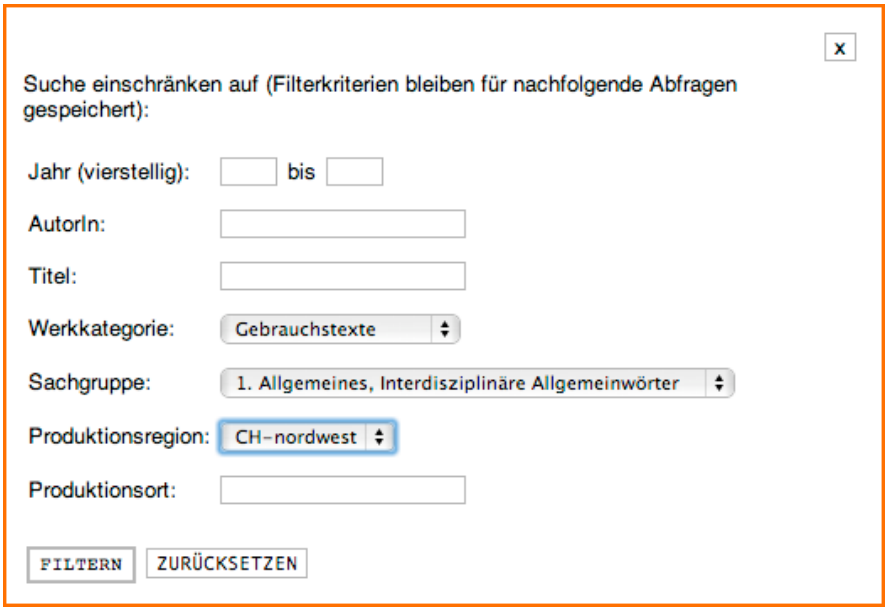


Abb. 11: Filter-Tool

5 Ausblick

Diese Beispiele zeigen, dass mit einem vergleichsweise kleinen, dafür aber gut strukturierten vielfältigen und ausgewogenen Korpus durchaus Fragen der Wortschatzentwicklung erforscht und dokumentiert werden können. Dabei dürfen die Resultate natürlich nicht verabsolutiert werden. Das erste Auftreten einer Neuschöpfung im Korpus liefert Hinweise auf den Entstehungszeitraum eines Wortes und ist nicht gleichzusetzen mit seinem tatsächlich ersten Auftreten in der Sprachgemeinschaft. Immerhin bietet sich das CHTK als ein nützliches Werkzeug an, mit dem solche Fragen wesentlich präziser angegangen werden können als bisher. Natürlich hat ein Korpus dieser Grösse auch Grenzen. Es gibt Wörter, die eindeutig zum Zentralrepertoire gehören und die im CHTK schlecht vertreten sind. Helvetismen wie *Januarloch* oder *Autoverlad* sucht man im CHTK vergeblich. Die vielversprechenden Ergebnisse, die mit den Abfragen zum zentralen Wortschatz bereits jetzt gewonnen werden können, zeigen aber, dass ein Ausbau mehr als lohneswert erscheint. Es bedürfte einer Pilotstudie, um genauere Aussagen darüber machen zu können, wie gross ein Korpus sein müsste, das möglichst keine Lücken hinsichtlich des Zentralrepertoires aufweist. Es ist zu hoffen, dass das CHTK in seiner aktuellen Grösse erst den ersten Schritt zu einem umfassenderen Korpus markiert.

Literatur

Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Berlin/New York: de Gruyter.

Ammon, Ulrich et al. (2004): *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin etc.: de Gruyter.

Berlin-Brandenburgische Akademie der Wissenschaften (ed.): *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts* (*DWDS*). www.dwds.de (Stand: September 2009).

Biber, Douglas (2007): "Representativeness in corpus design". In: Teubert, Wolfgang/ Krishnamurthy, Ramesh (ed.): *Corpus linguistics. Critical concepts in linguistics*. London, Routledge: 134–165. (Wiederabdruck eines Artikels in der Zeitschrift *Literary and Linguistic Computing*, 1993, 4/8: 243–257).

Biber, Douglas/Conrad, Susan/Reppen, Randi (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

Bubenhofer, Noah (2006): "Einführung in die Korpuslinguistik. Praktische Grundlagen und Werkzeuge." Elektronische Ressource, Zürich. www.bubenhofer.com/korpuslinguistik (Stand: September 2009).

Bubenhofer, Noah (2008): "Typischer Sprachgebrauch in Fachdiskursen. Corpus-driven-Analysen von großen Korpora". Referat an der internationalen Konferenz *Europhras* in Helsinki, 14. August 2008.

Buri, Emilie (2008): *Das SCHWEIZER TEXT KORPUS als empirische Basis für eine kulturlinguistische Untersuchung. Bedeutungswandel und Kultur im Schweizer Hochdeutschen des 20. Jahrhunderts*. Lizentiatsarbeit, Universität Basel.

Freie Universität Bozen/Europäische Akademie Bozen/Universität Innsbruck (eds.) (2005–): *Korpus Südtirol*. www.korpus-suedtirol.it/ (Stand: September 2009).

Gläser, Rosemarie (1990): *Fachtextsorten im Englischen*. Tübingen: Narr.

Klein, Wolfgang (2004): "Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts". In: Scharnhorst, Jürgen (ed.): *Sprachkultur und Lexikographie. Von der Forschung zur Nutzung von Wörterbüchern*. Frankfurt am Main etc., Lang: 281–309.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Niederhauser, Jürg et al. (eds.) (1999): *Wissenschaftssprache und Umgangssprache im Kontakt*. Frankfurt a. M.: Lang.

Österreichische Akademie der Wissenschaften (ed.): *AAC. Austrian Academy Corpus*. www.aac.ac.at/ (Stand: September 2009).

Rowland, Caroline. F./Fletcher, Sarah L./Freudenthal, Daniel (2008): "How big is big enough? Assessing the reliability of data from naturalistic samples". In: Behrens, Heike (ed.): *Corpora in language acquisition research. History, methods, perspectives*. Amsterdam/Philadelphia, Benjamins: 1–14.

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg: Winter.

Schnörch, Ulrich (2002): *Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und lexikografischen Aufarbeitung*. Tübingen: Narr.

Sinclair, John (1998): "Korpustypologie. Ein Klassifikationsrahmen". In: Teubert, Wolfgang (ed.): *Neologie und Korpus*. Tübingen, Narr: 111–128.

Teubert, Wolfgang/Čermáková, Anna (2007): *Corpus linguistics. A short introduction*. London: Continuum.