Masarykova univerzita
Pedagogická fakulta

Sprachkorpora
in Unterricht und Forschung DaF/DaZ

Tomáš Káňa

Brno 2014

Redaktion: Mag. phil. Sandra Reitbrecht
Rezension: prof. PhDr. Peter Ďurčo, CSc. (Bratislava/ Trnava)
Mag. Dr. Brigitte Sorger (Wien)

ISBN 978-80-210-6994-7

Inhaltsverzeichnis

Vorwort

Ein schneller Zugang zu Informationen ist heutzutage ebenso selbstverständlich wie die allgemeine Verbreitung der Computertechnik. Beides beeinflusst auch die heutige Sprachwissenschaft und Sprach(en)vermittlung. Tatsache bleibt jedoch, dass viele Studierende, (auch junge) Wissenschaftler/-innen und (lt. Erfahrungen aus Schulen) fast alle Lehrenden die Möglichkeiten der heutigen Instrumente für Forschung und Lehre nicht kennen oder kaum verwenden. Diese Tatsache kann unterschiedliche Gründe haben, die häufigsten aber sind das fehlende Wissen von der Existenz dieser Instrumente oder die Angst vor ihrer Komplexität, ihre (vermeintliche oder auch tatsächliche) Unübersichtlichkeit und die Aufwendigkeit ihrer Nutzung. Außer Kraft sollen beide Argumente mit diesem Buch gesetzt werden, welches jedoch mehrere Ziele anstrebt:

1) einen Überblick über die wichtigsten Begriffe der Korpuslinguistik, sowie der Typologie und Eigenschaften der größten Korpora zu geben. Das Auswahlverfahren beschränke ich auf diejenigen Instrumente, die
    a) dem Bereich DaF/DaZ nützlich sein könnten;
    b) kontrastiven Einblick unterstützen und ermöglichen;
    c) Schüler/-innen, Studierende und Lehrkräfte ansprechen können.
2) den Umgang mit Korpora einfach und transparent darzustellen und so die potentielle Angst vor den technisch oft aufwendigen Instrumenten abzubauen.
3) die Verbindung zwischen der Korpuslinguistik und DaF/DaZ herzustellen.
4) mit Studien zu einigen Phänomenen neue Erkenntnisse über die deutsche Sprache zu bringen. So leistet diese Monographie auch der (zum Teil) kontrastiven quantitativen Forschung der deutschen Sprache einen Beitrag.

In allen Punkten soll die Verbindung zu DaF/DaZ transparent und nachvollziehbar sein, auf Lerner- oder Fehlerkorpora wird jedoch nicht eingegangen. Alle Beispiele und Studien sollen auch als Anregungen für weitere und tiefere Forschungsfragen verstanden werden. Naturgemäß werden viele Aspekte ausgeblendet, einzelne Beispiele nicht bis ins letzte Detail interpretiert. Auch viele elektronische Instrumente werden nicht erwähnt, um die Übersichtlichkeit des Buches in Grenzen zu halten. In erster Linie sollen hier die vielfältigen Arbeitsmöglichkeiten mit elektronischen Instrumenten vorgestellt und die ersten Impulse für ihre Anwendung im Fremdsprachenunterricht initiiert werden.

Die technischen Anforderungen und das computertechnische Wissen der p.t. Leserschaft können minimalst sein: es reichen Basiskenntnisse im Umgang mit Computer und den gängigsten Programmen (Word, Excel, übliche Internetbrowser). Es ist keine Liebe zum PC erforderlich, sehr wohl aber die Liebe zur Sprache und eine gewisse Entdeckungslust.

Bemerkungen zur Buchgestaltung

Diese Publikation ist auch als ein „Kochbuch“ für die Korpusarbeit gedacht und gestaltet. Es muss zwar nicht chronologisch gelesen werden, für diejenigen, die mit der Korpusarbeit beginnen, ist es jedoch ratsam.

Nach der Erklärung der wichtigsten korpuslinguistischen Begriffe (Kap. 2) wird die Korpustypologie skizziert (Kap. 3). Im Kap. 4 (Korpora relevant für DaF/DaZ) werden konkrete Korpora vorgestellt: zuerst einige nationale Korpora von Sprachen, die im mitteleuropäischen Raum erfahrungsgemäß mehr Interessent/-innen haben als andere Sprachen (4.1), dann Korpora mit der deutschen Sprache (4.2) und zum Schluss das Projekt InterCorp, in dem teilweise alle Sprachen aus den Kap. 4.1 und 4.2 integriert werden. Während die nationalen Korpora anderer Sprachen im Kap. 4.1 nur kurz vorgestellt werden, wird in den beiden folgenden Kapiteln näher auf die Funktionen der Korpora eingegangen, um die Möglichkeiten ihrer Nutzung darzustellen. Zu den meisten Korpora, die an verschiedenen Institutionen auf dem deutschsprachigen Gebiet entstanden sind, gibt es gute Beschreibungen in Form der Manuale oder Hilfeportale (Hinweise sind im Text). Die Beschreibungen in diesem Buch sind „vorentlastend“ und bemühen sich, über einen einfachen Weg zu vernünftigen Rechercheergebnissen zu gelangen. Der Korpusmanager KonText (InterCorp) wird detailliert (im Kap. 4.2.4.2) beschrieben, weil zu diesem ein Manual auf Deutsch weder existiert noch geplant ist. Im Kap. 5 werden einige elektronische Instrumente beschrieben, die zwar keine Korpora sind, ihre Verwendung und Einsatz im Fremdsprachenunterricht kann jedoch ähnliche Hilfe leisten wie Korpora.
Das darauf folgende Kap. 6. (Studien) präsentiert konkrete Forschungsfragen und detaillierte Beschreibungen, wie man zu Ergebnissen kommt, und wie diese interpretiert werden können. Es ist ratsam, zuerst diese Grundinformationen über die Korpora im Kap. 4 zu lesen, im Internet nachzuschauen und dann den einzelnen Schritten im Kap. 6 zu folgen und womöglich gleichzeitig auch direkt im Internet durchzuführen.
Das vorletzte Kap. 7. (Statistiken) beinhaltet Listen mit statistischen Angaben über die deutsche Sprache, darauf (im Kap. 8) folgen Tabellen mit Kürzeln, Zeichen und Abfragemöglichkeiten, die für die Korpusarbeit erfahrungsgemäß nützlich sind.

Da sich die Schnittstellen einzelner Korpusmanager von einander gravierend unterscheiden, wurden im ganzen Buch folgende graphische Vereinheitlichungen durchgeführt, um die Bildhaftigkeit der einzelnen Schritte zu gewährleisten:

- Die Ikonen und/ oder Beschriftungen, wie sie auf dem Bildschirm erscheinen, werden in **Calibri** geschrieben.
- Ikonen zum Anklicken sind unterstrichen.
- Der nächste Schritt, Übergang zur nächsten Ikone oder zum nächsten Level wird durch einen Pfeil (→) markiert.
- Die Bestätigungstasten und Suchbuttons, die man anklicken muss, um zu den Ergebnissen (oder zum nächsten Schritt) zu gelangen, sind durch inverse Schrift (etwa Search) gekennzeichnet.
- Das konkrete Bild der Abfragen mit Sonderzeichen, Abfolge, Leerzeichen etc. wird in Calibri (10 Punkte) geschrieben.
- Sonderzeichen, die in den Abfragen vorkommen können, werden bei der Beschreibung der einzelnen Korpora erklärt und im Kap. 8 noch einmal zusammengefasst.
- Über Schrägstrich (/) werden im laufenden Text Synonyme oder Alternativen geschrieben.

Konkordanzen, syntagmatische Muster und andere Belege entsprechen dem Originalschriftbild aus dem Korpus - sie wurden nicht der jetzt geltenden Rechtschreibung angepasst, in einigen Belegen ist die Tokenisierung sichtbar (z.B. Leerzeichen vor einem Punkt oder Komma).
Bei Abfragen, die anspruchsvoller sind als einfache Abfragen nach einer Wortform, werden oft sog. reguläre Ausdrücke verwendet. Diese Ausdrücke (engl. regular expressions) sowie die Syntax ihrer Eingabe ins Suchfeld sind nur teilweise standardisiert. Aus diesem Grund sind die regulären Ausdrücke nicht zusammengefasst und werden bei den Beschreibungen der Abfragen beim jeweiligen Korpus angeführt. Falls sinnvoll wird auch ihre Bedeutung erklärt.

1. Einführung

Die Sprache ist für jede/-n Nutzer/-in ein selbstverständliches Kommunikationsmittel. Wir reden, schreiben, hören und lesen täglich unzählige Aussagen (Texte), über deren Form wir kaum nachdenken. Es gibt aber immer wieder Situationen, in denen wir uns die Frage stellen, ob diese oder jene Formulierung wohl richtig, üblich, normal oder akzeptabel, ja verständlich ist. Solche Fragen könnten (bzw. sollten) sich vielleicht öfter Schüler/-innen, Student/-innen und Lehrer/-innen stellen, die letzteren auch mit dem Gedanken im Hintergrund, wie sie das Wissen über die Sprache vermitteln und das Können ihrer Schüler/-innen weiterentwickeln.

Die Impulsfragen könnten auch wie folgt lauten: Wie finde ich schnell kurze Texte zu einem bestimmten Thema? Wie kann ich schnell zusätzliche Grammatikübungen angepasst an die konkrete Varietät des Deutschen zusammenstellen? Darüber hinaus stellen die Lernenden oft Fragen über die Sprache, auf die die Lehrer/-in vielleicht keine eindeutige Antwort (nicht einmal aus den Nachschlagewerken) parat hat. Solche Fragen sind in der Ganzheit der Sprache oft marginal, stellen jedoch einen Baustein im ganzen Mosaik des Systems der Sprache dar. Jede/-r Muttersprachler/-in kann zwar aus dem Stegreif Position beziehen, ob
(1) *Vergleich mit* oder *Vergleich zu* für ihn/sie „richtig klingt". Offen bleibt dabei allerdings, ob er/sie sich nicht womöglich vom eigenen Gefühl (Idiolekt) täuschen lässt?
Weitere aus dem „pädagogischen Leben" willkürlich gegriffene Fragen sind:
(2) Was ist im Deutschen gängiger: *sprechen von* oder *sprechen über*?
(3) Was entspricht den Verbindungen unter (1) und (2) im Englischen (Polnischen, Ungarischen, Türkischen)?
(4) Sagt man eher *der Pool* oder *das Pool*?
(5) Soll man jetzt eher *Peking* oder *Beijing/ Bombay* oder *Mumbai* sagen und schreiben?
(6) Was sagt die Norm zu (1) bis (5)?
(7) Was ist die Norm?

Die Fragen (1), (2), (4) sind typisch linguistische Fragen, die in der kommunikativen Sprach(en)vermittlung auf wenig Resonanz stoßen. Betrachtet man jedoch die Entwicklung in der Methodik der letzten Jahre, stellt man ein „Anzeichen für eine Weiterentwicklung von der kommunikativen zur kognitiven Wende [fest], wonach nicht mehr nur Sprachkönnen, als Ziel angestrebt wird, sondern das Wissen über die Sprache (wieder) in den Vordergrund rückt" (Rösch 2011: 89-91). Dies bedeutet jedoch keine Rückkehr zur Grammatik-Übersetzungsmethode, sondern, wie Rösch (2011: 90) schreibt, eher „die Erweiterung des Begriffes Grammatik zur *kommunikativen* Grammatik". Diesem Gedanken liegt der lexikalische Ansatz im Sinne des „lexical approach" (Lewis 1993) nahe, in dem vorgefertigte Sprachelemente, sog. „chunks", die Basisbausteine von Texten bilden. Somit bleibt der Text im Mittelpunkt der Sprachenvermittlung (das ist ja wichtig, denn die Kommunikation erfolgt in Texten), der Weg zur Textrezeption und -produktion wird jedoch mit solchen Bausteinen gepflastert, die über die Grenze eines Wortes hinausreichen, also mit Wortverbindungen und Phrasen. Diese Bausteine lassen sich mit den modernen (Sprach)Analyseinstrumenten relativ einfach erkunden, präsentieren und auch in Übungen umwandeln.

Viele linguistische, sprachenpolitische oder philosophische Fragen (wie die Beispiele unter (1) bis (7)) lassen sich erfolgreich mithilfe von Fachpublikationen beantworten, oft findet man die Antwort auch im Internet. Aber die einfache Frage danach, was gängiger ist, kann niemand ohne repräsentative Belege beantworten. In diesem Wort „gängig" steckt der Schlüssel zu Reflexionen über die Sprache an sich, aber auch über die zielführende Vermittlung von Sprache(n). Jede/-r Sprecher/-in hat zwar das Recht, die Sprache kreativ zu

verwenden, neue, ungewöhnliche Konstruktionen zu bilden, keine/r hat jedoch das Recht zu sagen „*X* sagt man nicht!“ oder sogar „*X* ist falsch!“, wenn es gleich mehrere Belege für *X* gibt. (Über diese streng präskriptive Phase ist hoffentlich jede demokratische Sprachenkommunität hinweg.)

Die Sprache als ein lebendes und lebendiges System ändert sich unabhängig von **einem** Individuum. Es ist ein System, das sich durch das Wiederholen der Elemente kennzeichnet. In anderen Worten: je öfter etwas (*X*) gesagt wird, desto größer ist die Wahrscheinlichkeit, dass dieses *X* von der Kommunität als richtig empfunden wird und (irgendwann auch) in normativen Werken standardisiert werden kann.

Die Betrachtungen einer Fremdsprache sollen sich auf diejenigen Aspekte konzentrieren, die allgemein gültig, verständlich und üblich sind. Diesem Ansatz folgend werden in diesem Buch Beispiele präsentiert, wie man sich die Arbeit mit der Erkundung der Sprache und der Erschließung einzelner Elemente erleichtern kann. Dazu kommen auch sprachkontrastive Einblicke auf einzelne Phänomene. So können auch Interferenzen oder Transfererscheinungen aufgedeckt werden.

1.1 Zugang zur Sprache über Suchmaschinen

Vieles lässt sich mit Suchmaschinen (z.B. Google, Altavista, Yahoo!) auch über die Sprache(n) finden: interessante Artikel, Wörterbücher, Diskussionsforen über sprachliche Elemente, Unterrichtsmaterialien u.a., wie es auch die folgenden Beispiele verdeutlichen.

Einmal wollte ich wissen, ob die Wortzusammensetzung *Korruptions-Untersuchungsausschuss* auch außerhalb der österreichischen Realität existiert: am 16.4.2012 fanden sich unter Google.com 27.000 (!) Nachweise . Aus den Homepage-Adressen der ersten Seiten ließ sich erahnen, dass es sich hauptsächlich um österreichische Medien und Portale handelte (Domäne: .at).
Dieselbe Abfrage stellte ich am 9.1.2014 an Google.com. Die Suche ergab „nur“ über 22.400 „Treffer“, bei yahoo.at (erweitert auf die ganze Welt) lediglich 2.620 Ergebnisse. Leider ist aus diesen Angaben nicht ersichtlich, ob das Wort (noch) im Umlauf ist, wie sich die Frequenz veränderte, in welchen Texten und wie oft es vorkommt.
Aus der schrumpfenden Anzahl der Belege lässt sich nur erahnen, dass der Begriff nicht mehr so gebräuchlich ist. Viele Artikel sind offensichtlich aus dem Internet verschwunden. Es lässt sich auch nicht überprüfen, ob die Anzahl von 27.000 im Jahr 2012 tatsächlich gestimmt hat, ob ich mich vielleicht um eine Null nicht vertippt habe.

In einem Sprachkorpus (DeReKo) war am 16.4.2012 kein einziger Treffer. Zwei Jahre später, (9.1.2014) fand ich zwar nur 37 Treffer, aber auch aus diesen Daten lassen sich bereits Informationen ableiten, die (anhand der geringen Belegmenge) mit etwas Vorsicht Folgendes aussagen:

a) Die Wortkopplung *Korruptions-Untersuchungsausschuss* kommt fast ausschließlich in österreichischen Texten vor.
b) Weniger überraschend ist, dass es sich um Zeitungstexte über Politik handelt.
c) Am häufigsten wurde der Wortkomplex im März 2012 verwendet.

Es sind zwar keine überraschende Ergebnisse, es sind aber Beweise über das gesuchte Wortgeschöpf.

Als ein Problempunkt tauchte in einem DaF-Seminar die Frage nach der korrekten Präposition auf: Ist die Verbindung *im Vergleich mit* (*etwas*) oder *im Vergleich zu* (*etwas*) richtig? Die übliche Praxis sagt: man sucht auf gut Glück, was es gibt: ins Google-Suchfeld schreibt man „Vergleich mit oder zu" und bekommt Links zu mehr oder weniger seriösen Artikeln oder Foren.

Im Forum Deutsch als Fremdsprache (1996-2012) löste man die Problematik folgend:
Frage:

> **Im Vergleich mit/oder zu?**
> geschrieben von: **Hussein** ()
> Datum: 18. Mai 2009 19:25
> *Hallo zusammen,*
> *man hat mich korrigiert, in dem man den Satz "Im Vergleich mit..." durch den Satz "Im Vergleich zu…" ersetzt hat. Könntet ihr mir mal sagen, was der Unterschied dazwischen ist. Im Wörterbuch stehen aber beide Möglichkeiten.*

Antwort:

> Re: Im Vergleich mit/oder zu?
> geschrieben von: oberhaenslir ()
> Datum: 18. Mai 2009 23:02
> *Standarddeutsch: etwas mit etwas vergleichen; jemanden mit jemandem vergleichen*
> *Korrekt ist also 'im Vergleich mit':*
> *"Schweizer Hochschulen halten im Vergleich mit Deutschland und Österreich gut mit."*
> *Falsch ist 'im Vergleich zu':*
> *"Wie schneidet ein heutiges Gerät im Vergleich zu einem Geschirrspüler vor zwanzig Jahren beim Wasserverbrauch ab?"*

Andere Forumsteilnehmer versuchen die Antwort zu relativieren:

> Re: Im Vergleich mit/oder zu?
> geschrieben von: Franziska ()
> Datum: 19. Mai 2009 16:51
> *Ich glaube, heutzutage wird im Alltag mehr "Im Vergleich zu" verwendet - das könnte ein Grund für die Korrektur sein, Hussein. Aber sicher sind beide richtig. Dieses "zu" drückt wohl mehr die Richtung aus, in die der Vergleich abzielt. Es klingt ein bisschen energischer. "Mit" klingt gleichwertiger, freundlicher.*
> *"Im Vergleich zu mir hat er keine Ahnung, ha!"*
> *"Im Vergleich mit dieser Teesorte schmeckt die andere etwas lieblicher."*
> *Aber das ist nur so ein Gefühl - bevor wieder Jeros obligatorisches "Hm" kommt.*

Es herrscht also Unsicherheit. Ob Hussein jetzt weiß, was er das nächste Mal schreiben soll, ist zu bezweifeln. Die Frage bleibt unbefriedigend beantwortet.
Die Google-Suche nach den genauen Verbindungen „*im Vergleich mit*" und „*im Vergleich zu*" zeigt ziemlich deutlich, dass die Verbindung mit der Präposition *zu* überwiegt (187 Mio. Suchergebnisse zu 51 Mio. Suchergebnissen mit der Verbindung *im Vergleich mit*). Einige von den Ergebnissen repräsentieren Artikel, Foren und Polemiken zu dieser Problematik (wie hier oben angeführt), für das wahre Bild der Verbindung sind sie also irrelevant. Texte, in denen diese Verbindungen in einem natürlichen Kontext vorkommen, sind schwer herauszufiltern. Dabei lässt sich das Problem anhand von einigen realen Textpassagen induktiv ziemlich schnell und zufriedenstellend beantworten:

Sie wiegen ja viele Tonnen , und	im Vergleich	**zu** ihrer Größe ist das Gehirn relativ klein.
Das Wasser war kalt	im Vergleich	**zu** der warmen Luft
Es war sogar ganz außerordentlich klein	im Vergleich	**zu** anderen Ländern , wie zum Beispiel Deutschland
wird festgelegt , und dadurch	im Vergleich	**zu** allen anderen ,
- doch	im Vergleich	**zu** der Macht , die
Die glauben doch , daß	im Vergleich	**zu** ihrer Arbeit alles andere völlig bedeutungslos ist

Es waren freie und glückliche Jahre	im Vergleich	**zu** den Zeiten , die mich im Kloster erwarteten ,
Ein Übel , das nichts ist	im Vergleich	**mit** dem , was dich in der Hölle erwartet ,
Aber dieser Tick war gar nichts	im Vergleich	**zum** Verhalten seines Vaters , der sich…
manche der Aussagen des Thomas Morus wirken	im Vergleich	**mit** diesem 20. Jahrhundert frappant

Aus diesen zehn (zufällig aus 1.533 ausgewählten) Beispielen aus dem Korpus InterCorp ist auf den ersten Blick ersichtlich, dass
1) kein Beleg aus einer linguistischen Polemik über die Konkurrenzformen stammt;
2) beide Präpositionen gebraucht werden;
3) mehrheitlich *zu* verwendet wird. Das Verhältnis ist nicht 2:8, wie es hier den Anschein erweckend dargestellt ist, sondern etwa 1:100. Die Verbindung *im Vergleich mit* ist also eher selten.

Es ist unumstritten, dass die Suchmaschinen wertvolle Daten liefern, aus denen man sich einen ersten Überblick verschaffen kann: sie sind blitzschnell, liefern eine große Menge von Ergebnissen, darüber hinaus sind die Daten höchstaktuell – entsprechen dem momentanen Zustand im Internet. Aus den Ergebnissen der Internetsuchen kann man bestimmt sehr gut die erste Orientierung auch über sprachliche Phänomene gewinnen[1]. Zumindest die simple Verifizierung, ob das Abgefragte existiert oder nicht, kann wertvoll sein. Ob man aus Internetbelegen Beispiele zum Sprachgebrauch präsentieren, Übungen effektiv zusammenstellen und überhaupt die Sprache mit bloßen Internetrecherchen systematisch erkunden kann, ist jedoch ernsthaft anzuzweifeln.

1.2 Korpustools in gängigen Programmen

Ein „Korpus“ kann jeder durchschnittliche PC User erstellen und so die Arbeit mit den Daten üben.
Die Vorstellung von den elementaren Korpusfunktionen verschafft man sich über einige Tools heute gängiger Computer-Programme:

MS-Word hat beispielsweise diese sprachlichen Tools:
- auf der Registerkarte Überprüfen in der Gruppe Dokumentprüfung → Wörterzählen gibt es statistische Angaben zum Text: Seiten, Absätze, Zeilen, Zeichen (mit oder ohne Leerzeichen) Rechtschreib- und Grammatikkorrektor
- Suchen von Wort(teilen) über die Tastenkombination Strg+F
- (Übersetzungsdienste nur online)

Google-Suche verfügt unter Erweiterte Suche über diese Suchmöglichkeiten:
Suche nach
- all diesen Wörtern: Diese Funktion ermöglicht teilweise lemmatisierte Suche, d.h. die Abfrage kalt UND kochen ergibt auch kalten (Tee), kalte (Küche) etc. und weiter im Text kochen.
- der exakten Phrase
- einem/ keinem dieser Wörter

[1] Die Validität der Ergebnisse lässt sich dann etwa mit dem Informationswert der Wikipedia-Artikel vergleichen: Vieles entspricht der Realität, ist aber manchmal einseitig, nicht genügend belegt oder zu oberflächlich.

Diese Tools sind hilfreich für die Suche nach dem gewünschten Thema. Was die Ermittlung der Sprache anlagt, haben die Programme allerdings ziemlich eingeschränkte Möglichkeiten. Dem ist es so aus dem einfachen Grund: sie sind nicht primär für Fragen über die Sprache per se bestimmt.
Die wichtigsten Unterschiede zwischen einer Suchmaschine und einem Korpus bezüglich der Spracherkundung sind in der folgenden Tabelle zu sehen.

	Suchmaschine	**Korpus**
Geschwindigkeit	schnell	oft langsam
Benutzerfreundlichkeit	einfach, intuitiv	oft kompliziert, ohne Einleitung kaum verständlich
Daten(Text)menge	riesig	oft (sehr) eingeschränkt
Wesen der Sprache	nur geschrieben	oft nur geschrieben, aber auch gesprochen
Sprache der Texte	eingeschränkt eindeutig	eindeutig
Originalsprache d. Texte	kaum feststellbar	eindeutig feststellbar
Zeit der Verfassung	kaum feststellbar	eindeutig feststellbar
Textfunktion	keine Angaben	(eindeutig) zuordenbar
Verfasser/-in(nen)	schwer feststellbar	relativ eindeutig
Strukturiertes Suchen im Text (in Texten)[2]	nicht möglich	möglich
Suche nach Wortteilen	eingeschränkt möglich	möglich
Statistische Angaben	fast keine (nur Anzahl der Ergebnisse)	vielfältig

Tab. 1: Vergleich einiger Eigenschaften: Suchmaschine - Korpus

Die Unterschiede werden auch aus den Bestimmungen der Instrumente klar: eine Suche im Internet ist der Suche in einem Bibliothekskatalog oder einer Enzyklopädie ähnlich. Dort findet man jedoch (fast) keine sprachlichen, grammatikalischen Informationen über die gesuchten Begriffe. Dafür gibt es andere Quellen: Wörterbücher und Grammatiken. Korpora hingegen sind keine Suchmaschinen nach Themen, sondern nach sprachlichen Erscheinungen, die dann interpretiert werden müssen.

Beide Instrumente können sich in der Sprachforschung und -vermittlung gut ergänzen: die Internetsuchmaschinen liefern „frische“ Daten über den Stand der Internetsprache, Korpora können wiederum gut auch über (vergangene) gesellschaftliche Themen informieren: Schaut man im Korpus[3] nach dem Wort *Impeachment* nach, stellt man sofort fest, womit sich die Welt (unter anderen wichtigen Ereignissen) in den Jahren 1998 und 1999 befasst hat. Google bietet heutzutage (nur) Hinweise zu Seiten mit Erklärungen dieses Wortes.

[2] Z.B. Suche nach zwei Wörtern in einem Satz, oder Wortanfang/-ende, nach grammatikalischen Angaben (z.B. nur Adjektive, die auf *-el* enden).

[3] DeReKo: W - Archiv der geschriebenen Sprache.

2. Begriffsbestimmungen

Für ein reibungsloses Verständnis des Inhalts dieses Buches werden hier die wichtigsten Begriffe definiert und erklärt, wie sie in der Korpuslinguistik üblicherweise verstanden werden. In diesem Kapitel werden sie thematisch geordnet: von Korpus und Sprachkorpus ausgehend bis zu einigen korpustechnischen Begriffen wie Parser oder Alignment, die für die Recherche zwar entbehrlich sind, in Korpusmanuals oder in der Fachliteratur jedoch oft vorkommen. Alphabetisch werden die einzelnen Begriffe noch einmal (mit einer kurzen Definition) im Register angeführt.

2.1 Korpus und Sprachkorpus

Ein **Korpus** ist lateinisch „Körper". Selbst das deutsche Wort *Körper* ist eine sprachliche Entlehnung aus dem Lateinischen und fungiert laut Kluge (2002: 530) in der deutschen Sprache seit dem 13. Jh.:

> „…mhd. korper, körper, fnhd. auch körpel mit Dissimilierung des zweiten r Entlehnung. Entlehnt aus l. corpus (-poris) n. "Leib" (der spätere Umlaut ist nicht ausreichend erklärt). Ersetzt die älteren Wörter Leib und Leiche."

In der heutigen deutschen Sprache gehen mehrere Lexeme auf diese gemeinsame Wurzel (Etymon) zurück: *Körper*, *Korps*, *Korpuskel*, *korpulent*, und natürlich auch *Korpus* oder *Corpus*. Laut DUDEN (1996: 885-886 und 2006 [CD-ROM]) ist **der** Korpus die Bezeichnung des menschlichen Körpers) und **das** Korpus eine Sammlung bzw. der Körper eines Instruments:

> [1]**Korpus**, der; -, -se [lat. corpus, ↑Körper]: **1.** (ugs. schrezh.) *menschlicher Körper* … **2.** (bild. Kunst) *Christusfigur am Kruzifix*. **3.** <o. Pl.> (Fachspr.) (bei Möbeln) *das massive, die eigentliche Gestalt ausmachende Teil ohne die Einsatzteile* … **4.** (schweiz.) *Ladentisch; [Büro]möbel mit Fächern* …;
> [2]**Korpus**, Corpus, das; -, Korpora bzw. Corpora [lat. corpus, = Gesamtwerk, Sammlung…]: **1.** (Sprachw.) *Sammlung einer begrenzten Anzahl von Texten, Äußerungen o. Ä. als Grundlage für sprachwissenschaftliche Untersuchungen.* **2.** <heute meist: der; o. Pl.> *Klangkörper bes. eines Saiteninstruments*;
> [3]**Korpus**, die; (Druckw.) *Schriftgrad von 10 Punkt; Garmond*.

Im Kontrast zu den angeführten Angaben bringt eine Recherche im größten deutschen Korpus/ Corpus Ergebnisse, die auf den folgenden Gebrauchsusus hindeuten:

- Das Schriftbild mit „C" kommt in modernen deutschen Texten der letzten 20 Jahre nur in einer lateinischen Verbindung vor: z.B. *Corpus delicti*, *Corpus Christi*, *Ave verum corpus*, *Corpus iuris Civilis*.
- Bis auf *Corpus Christi* ist das Lexem immer nur sächlich. In der Verbindung *Corpus Christi*, bzw. nur elliptisch *Corpus* ist es ausschließlich männlich[4].
- *Korpus* mit *K* geschrieben ist fast immer maskulin, egal ob es sich um einen menschlichen Körper, das Geschöpf eines Instrumentes oder eine Sammlung handelt (alle folgenden Belege sind aus dem DeReKo, W – Archiv der geschriebenen Sprache):

  > Die rhythmisch pointierte Sprache der eigenwilligen Geige kommt am eindrucksvollsten bei ihren perkussiven Elementen zur Entfaltung, wenn er die Saiten zupft und **den Korpus** als Trommel zusammenwirken läßt.
  >
  > Eine repräsentative und hinreichend große Menge an verfügbaren schriftlichen Quellen, genannt **der Korpus**, aus einer Sprache wird als Grundlage für die Auswertung verwendet.

[4] Ein Beleg aus einem Dutzend im DeReKo: *Das neue, fast drei Meter hohe Kreuz ist aus Holz und trägt keinen Corpus.*

Deswegen wird auch in manchen Kinos von Wolfram Weber weiter ein analoger Zelluloid-Projektor stehen. Sogar wenn **der** gesamte **Korpus** der Filmgeschichte einst digitalisiert sein sollte.

- Ein seltener Fall ist das neutrale Genus von *Korpus* im Sinne einer Sammlung, wie das folgende Textbeispiel zeigt:

  Kurt Ostbahn. […] Nach drei überfüllten Abenden en suite hängt der echteste aller Wiener noch zwei Auftritte im "Zelt" an. Der Doppler nennt sich das Ereignis, bei dem **das** gesamte historische **Korpus** des Kurti-Schaffens - auf zwei Konzerte verteilt - ausgebreitet werden wird, natürlich ohne auch nur eine Nummer doppelt zu spielen.

Der nächste interessante Punkt im oben genannten Lexikoneintrag (Duden Universalwörterbuch 1996, und in allen späteren Versionen) stellt die Bedeutungserklärung dieses Lexems dar. An der ersten Stelle (dies bedeutet wohl die häufigste Verwendung) steht:

1. (ugs. schrezh.) *menschlicher Körper* …

Sieht man das typische Umfeld[5] von *Korpus* in der geschriebenen deutschen Sprache an, findet man keinerlei Hinweise auf umgangssprachliche oder scherzhafte Elemente:

ein gusseiserner **Korpus**
Wegekreuz … mit|bezeichnet … **Korpus** bezeichnet …
der|die Saiten und|auf den **Korpus**
spätbarockes Friedhofskreuz mit **Korpus** … 1755
Friedhofskreuz mit **Korpus**
einen … hohlen […] **Korpus** der
mit hohlem **Korpus**
ein Kruzifix [mit …] **Korpus**
des|Hohler **Korpus** … die|der … Zargen mit Tonabnehmer
mit massivem […] **Korpus**

Trotz der Diskrepanz zwischen der Norm und dem realen Gebrauch wird in dieser Monographie die pragmatische Lösung verfolgt, nämlich konsequent der Terminus ***das*** *Korpus* verwendet, weil in der korpuslinguistischen Fachliteratur das neutrale Geschlecht und die Schreibweise mit *K* immer noch überwiegen. Mit *Korpus* ist hier nur das *Sprachkorpus* gemeint.

Der heutige Begriff **Sprachkorpus** bezeichnet eine elektronische Sammlung/ Datenbank natürlicher Texte, die (meist) in voller Länge gespeichert wurden[6]. Diese Datenbank ist strukturiert und mit einem **Korpusmanager** versehen. Diese zwei Eigenschaften heben ein Korpus von einer einfachen Textdatenbank oder von einer elektronischen Bibliothek ab. (Dazu auch Káňa/ Pelouškova 2005 und McEnery/ Wilson 2001: 30). Als ein Sprachkorpus kann nicht jede (wenn auch wertvolle) elektronische Datenbank bezeichnet werden. Die Mindestanforderungen an die Suchmöglichkeiten in einem Sprachkorpus sind: Abfragen nach Wörtern und Wortteilen, statistische Angaben über die Ergebnisse, Angaben über die Größe und Zusammenstellung des Korpus. Ein Sprachkorpus ist auch kein Wörterbuch, reflektierte User können es jedoch als solches verwenden.
Ein Sprachkorpus ist also eine elektronische Textdatenbank, in der man effektiv nach sprachlichen Phänomenen suchen kann.

[5] Sog. „syntagmatische Muster“ (Belica 1995).
[6] Die Vorreiter elektronischer Korpora waren Kartotheken, in denen nur Textsequenzen eingetragen wurden.

Sprachkorpora hat es schon lange im vorelektronischen Zeitalter gegeben. Unentbehrlich waren sie für die Erstellungen seriöser Wörterbücher, wie im Vorwort zu Band 1 des Grimmschen DWB (1854-1961: XXXVII) angeführt wird:

> „Wörter verlangen beispiele, die beispiele gewähr, ohne welche ihre beste kraft verloren gienge. wie könnten stellen (loci) heiszen, deren stelle ungenannt bliebe? der name ihres urhebers reicht nicht aus, sie müssen aufgeschlagen werden können; aus der leichtigkeit dieses nachschlagens entspringt ein groszer reiz, denn wie genau auch die belege ausgehoben seien, der leser hat nicht selten das bedürfnis sie in ihrem vollständigeren zusammenhang einzusehen: indem er weiter vordringt, findet er dicht neben den beigebrachten ausdrücken noch etwas anderes, unmitgetheilt gebliebenes, wodurch ihm das verständnis vollends erschlossen wird. auch in der classischen philologie ist es hergebracht die quelle anzuführen, aus der entnommen wurde. unbelegte citate sind unordentlich zusammengerafte, unbeglaubigte, unbeeidete zeugen.“

Prinzipiell wurde jede Sammlung von Texten, die für Zwecke sprachlicher Untersuchungen zusammengestellt wurde, als ein Sprachkorpus bezeichnet. In der heutigen üblichen Verwendung beschränkt sich der Begriff auf elektronische Sprachkorpora.

Ein **Korpusmanager** („Korpussuchmaschine“) ermöglicht eine effektive Suche nach sprachlichen Elementen. Korpusmanager sind aufwendige Text-Analyse-Systeme, die gezielte, auch kompliziertere linguistische Recherchen ermöglichen.

Die Suche erfolgt in Form einer **Abfrage**. Allerdings ist die Sprache des Korpusmanagers anders als die „normale“ Sprache. Deswegen muss die Abfrage in die Korpussprache „übersetzt“ werden. Die Sprache heißt *corpus query language* (**CQL**).

CQL (Corpus Query Language) ist eine spezielle „Sprache“, in der man mit dem Korpus-Manager (z.B. CQP – Corpus Query Processor) kommuniziert, d.h. die Abfrage erstellt. Für die Abfragestellung gibt es (je nach Korpusmanager) entweder die sog. „graphische“ Eingabe oder die „zeilenorientierte[7]“ Eingabe der Abfrage. Die graphische Eingabe wird durch Klicken auf Ikonen, die zur Auswahl stehen, zusammengestellt. Die zeilenorientierte Eingabe der Abfrage wird direkt ins Suchfeld getippt.

Die Formulierung der **Abfrage**n ist fast in jedem Korpusmanager anders. Ohne Basiswissen über das Korpus, den Korpusmanager und dessen Eigenschaften ist eine kompliziertere Abfragestellung oft schwierig, manchmal sogar unmöglich. Aus diesen Gründen muss jeder Nutzer die Manuale (falls sie vorhanden sind) der Korpusmanager studieren, die Eigenschaften kennen oder einfach probieren, was geht. (Diesen langwierigen Prozess soll dieses Buch erleichtern und gleichzeitig dafür plädieren, dass das Bassiswissen über diese Instrumente in jedem Sprach(en)unterrricht verbreitet wird.)

Der Begriff **Abfrage** beinhaltet 1) die Einstellung der Suche, 2) die Einstellung der Ergebnispräsentation und 3) die Formulierung der Frage an den Korpusmanager im Suchfeld (**Suchfeldeingabe**).

Die **Suchfeldeingabe** (oder nur **Eingabe**) ist eine Kombination von Graphemen, Zeichen und anderen Elementen, die ins Suchfeld des Korpusmanagers eingegeben werden muss, um die gewünschte sprachliche Erscheinung aus den Korpustexten abrufen zu können.

[7] Begriff, der in COSMAS II verwendet wird.

Übersicht der Abfragemöglichkeiten
Die Möglichkeiten der Abfrage sind grundsätzlich diese:
Suche nach

1) einzelnen Wörtern bzw. Zeichenketten, die ein Bestandteil eines Wortes sind
2) Grundformen
3) Wortkombinationen
4) Kombinationen (von Wörtern, Zeichenketten etc.) mit Abstand voneinander
5) morphosyntaktischen Kategorien (in Kombination mit Abstand)
6) Satzelementen

Die Möglichkeiten der Suchabfragen variieren nach Korpustyp. Als Ergebnis einer Korpusabfrage erscheinen auf dem Bildschirm **Konkordanzen**.

Die **Konkordanzen** in der Korpuslinguistik[8] sind Ergebnisse einer Suchanfrage. Es sind **Belege** aus den Korpustexten. Sie erscheinen in Form einer Konkordanzzeile (Abb. 1), in der Mitte steht das gesuchte Element vom umliegenden Text graphisch hervorgehoben (in der Abb. 1 Schaf), genannt **KWIC** (steht für **K**ey **W**ord **I**n **C**ontext), oft wird es auch **Node** oder **Treffer** genannt. Rechts und links auf derselben Zeile sind Wörter aus der Umgebung im Text:

Wenn sie oben am Hügel standen , sahen sie	**Schafe**	und Ziegen auf einem grünen Rasen springen .
Der Pater begann zu glauben, das gute	**Schaf**	aus des Herren Herde sei nicht richtig im Kopf
, ich bin das schwarze	**Schaf**	der Familie , ein Windhund .
(…) erfuhr Esteban Trueba von dem schwarzen	**Schaf**	in seiner Familie nur durch den (…) Briefwechsel
Gib mir Kraft , die Sünderin zu bessern , das verirrte	**Schaf**	in deine Herde zurückzubringen !

Abb. 1: Konkordanzen zur Abfrage **Schaf** (InterCorp_de)

Zu diesen Konkordanzen gibt es auch Angaben über die Texte, aus denen sie stammen. Sie werden jedoch (je nach Korpusmanager) unterschiedlich ausführlich angegeben.

Zu den ersten Texten, die als „Sprachkorpora“ verwendet wurden, zählt die Bibel. Um in der Heiligen Schrift besser suchen zu können, wurden „Konkordanzen“ erstellt: ein Verzeichnis der Wörter mit Kennzeichnung der Stelle, wo sie vorkommen. Die erste Konkordanz entstand zur Vulgata Anfang des 13. Jahrhunderts, später folgten Konkordanzen zu den Bibel-Übersetzungen. Heute ist die Suche in mehreren deutschen Übersetzungen auch online möglich - beispielsweise über Bibel Online (2014), wie die Abb. 2 zeigt:

1.) Und ich will meiner Herde helfen, daß sie nicht mehr sollen zum Raub werden, und will richten zwischen **Schaf** und **Schaf**. (Hesekiel 34.22)
2.) Aber zu euch, meine Herde, spricht der Herr, HERR also: Siehe, ich will richten zwischen **Schaf** und **Schaf** und zwischen Widdern und Böcken. (Hesekiel 34.17)
3.) Wenn jemand einen Ochsen oder ein **Schaf** stiehlt und schlachtet's oder verkauft's, der soll fünf Ochsen für einen Ochsen wiedergeben und vier **Schafe** für ein **Schaf**. (2. Mose 21.37)
4.) Es sei ein Ochs oder **Schaf**, so soll man's nicht mit seinem Jungen auf einen Tag schlachten. (3. Mose 22.28)

Abb. 2: Konkordanzen der Abfrage **Schaf**, Bibel Online (Übersetzung: M. Luther, Ausgabe 1912).

Diese modernen Instrumente verwenden für die Suche spezielle **Konkordanzprogramm**e. Es sind elektronische Instrumente für die Suche nach Wortformen in beliebigen elektronischen

[8] Das Wort *Konkordanz* hat mehrere Bedeutungen. In der Schweiz wird mit *Konkordanz* u.a. die Koalitionsregierung bezeichnet.

Texten. Auch die Basis jedes Korpusmanagers bildet ein Konkordanzprogramm. Das Ergebnis der Suche mit einem Konkordanzprogramm sind Konkordanzen in Konkordanzzeilen. (Im Internet findet man unter den Begriffen *Konkordanzprogramm* oder *concordancer* viele Konkordanzprogramme mit unterschiedlich komplexer Bedienung und diversen Funktionen.)

Bei der Auswertung der Konkordanzen/ Treffer sind weitere Funktionen der Korpusmanager wichtig, die die Interpretation der gewonnenen Daten erleichtern oder gar erst ermöglichen. Die üblichsten Funktionen, die in den meisten Korpora vorkommen, sind: **Filter**, **Frequenz-** und **Kollokationsanalysen**.

Der **Filter** ermöglicht die Belege zu sortieren. Prinzipiell kann man positive und negative Filter einsetzen.

In jedem Korpus gibt es mehrere Angaben zu den **Frequenzen** (Häufigkeitsmaße) von KWIC. Grundsätzlich existieren in jedem Korpus zwei Frequenzgruppen:

1) Frequenzangaben zum KWIC:

absolute Frequenz: exakte Anzahl des Vorkommen (der Treffer) im Korpus;

relative Frequenz: Angabe über das Vorkommen bezogen auf eine Menge (z.B. die Gesamtgröße des Korpus oder das absolute Vorkommen des gesuchten Elements). In vielen Korpora wird der (hypothetische) Richtwert *pro eine Million Worte* (p.m.W.)/ *instances per million* (i.p.m.) berechnet.

Häufigkeitsklassen: Werte von 0 bis n. Die Häufigkeitsklassen werden oft auf das Vorkommen des am meisten frequentierten Wortes in der Sprache bezogen: je kleiner die Ziffer, desto wahrscheinlicher ist das Vorkommen des Wortes im (beliebigen) Text. Das Zentrum des Wortschatzes (allgemein verständliche, hochfrequentierte Wörter) fällt in die Häufigkeitsklassen 0 bis 13 (insgesamt 6.700 Grundformen). Wörter mit der Häufigkeitsklasse höher als 18 kann man als äußerst selten bezeichnen.
Die Häufigkeitsklasse 0 haben im Deutschen die Wörter *der*, *die* und *und* (vgl. DeReWo 2009).

2) Frequenzangaben des KWIC zu anderen Wörtern:
Kookkurrenzen, bzw. **Kollokationen**.

Obwohl die **Kookkurrenz** als ein gemeinsames Erscheinen (ohne Betonung der semantischen Bindung) verstanden wird und die **Kollokation** als eine semantische Verbindung mehrerer Elemente (vgl. auch Ďurčo 1994: 16) definiert werden, erscheinen in der korpuslinguistischen Literatur beide Begriffe mehr oder weniger als Synonyme.
Die Berechnung der Kookkurrenzen/Kollokationen zeigt das typische Umfeld des KWIC. Ein übersichtliches Tutorial zur Kookkurrenzanalyse ist auch auf der Homepage des IDS-Mannheim (siehe Belica 1995) zu finden.
Die Kollokationen werden nach unterschiedlichen (statistischen) Formeln berechnet, die den Kollokationswert/ das Signifikanzmaß ergeben. Dieser Wert (jeweils eine Nummer) kann man sich als ein Signal für die Wahrscheinlichkeit des gemeinsamen Vorkommens vom KIWC und seinem Kollokationspartner (Kollokator) vorstellen. Je nach den Variablen, die in diese Berechnung einbezogen werden (z.B. Größe des Korpus, Häufigkeit des KWICs und/ oder

des Kollokationspartners, Anzahl der Dokumente im Korpus etc.), und nach dem Algorithmus, der die Signifikanz berechnet, werden auch die Signifikanzmaße benannt: *mutual information score* (**MI-score**), *test score* (**T-score**), *log likelihood ratio* (**LLR**) um nur die üblichsten zu nennen. Die Werte werden bei den Kookkurrenz-/Kollokationsanalysen in jedem Korpus automatisch mitgeliefert (siehe auch die Beschreibung von DWDS, DeReKo und InterCorp).

Mi-Score (mutual information) ist ein Wert, der die Wahrscheinlichkeit der Kookkurrenz von KWIC und dem Kookkurrenzpartner (Kollokator) anzeigt. Ist der Wert höher als 10.000, handelt es sich höchstwahrscheinlich um feste Verbindungen (bzw. Namen, die sich im Korpus wiederholen).

T-Score (test score) „relativiert" den Mi-score, zeigt Partner, die weniger häufig, jedoch immer noch signifikant vorkommen. Die Reihung nach dem T-Score zeigt (je nach Einstellung) deutlich auch das „grammatikalische Umfeld" vom KWIC: Funktionswörter, Präpositionen, Satzzeichen.

LLR (log likelyhood ratio) bedeutet „Interner Wert für die ermittelte Stärke der lexikalischen Kohäsion" (Belica 1995). Es ist immer eine positive Nummer, sie kann beliebig hoch sein (je frequentierter beide Kollokationspartner sind, desto höher ist sie). Die Werte der Kollokatoren sinken bis 0. Dieses Signifikanzmaß ist ein guter Kompromiss zwischen dem Mi-Score und dem T-Score und wird in den meisten Korpora berechnet.

logDice ist eine Berechnung, die für Lexikographen empfohlen wird (Rychlý 2008: 6), da sie typische (aber nicht unbedingt häufige) Partner aufdeckt, untypische jedoch ausblendet. Aus diesem Grund könnte sie auch für den DaF/DaZ-Unterricht verwendet werden. Leider werden diese Werte nicht von jedem Korpusmanager[9] berechnet. Die Vorteile dieses Maßes listet Rychlý (2008: 9) folgend auf:

- Der maximale Wert kann 14 sein, üblicherweise ist er unter 10.
- Der Wert 0 (oder negativer Wert) heißt (absolut) keine Signifikanz.
- Das Maß ist von der Größe des Korpus unabhängig.

Im beliebig großen (Sub)Korpus kann man sich also sicher sein, dass die Kollokatoren mit dem Wert über 10 sehr häufig mit dem KWIC vorkommen (d.h. diese sollten die Lerner/ Lernerinnen unbedingt beherrschen), dafür können die nahe 0 ausgeblendet werden. Als Beispiel kann man sich die typischen Kollokatoren zur Verbform *fragst* anschauen:

	Filter		Freq	T-score	MI	logDice
1.	p/n	Wieso	5	2.235	10.670	7.002
2.	p/n	Warum	23	4.792	10.212	6.742
3.	p/n	du	155	12.439	10.116	6.681
4.	p/n	Ernst	4	1.998	10.331	6.666
5.	p/n	Du	37	6.076	9.923	6.473
6.	p/n	zuviel	3	1.730	10.125	6.426
7.	p/n	warum	12	3.457	9.027	5.563
8.	p/n	dich	20	4.463	8.965	5.516

Abb. 3: Kollokationspartner zur Abfrage *fragst* (InterCorp_de)

[9] DWDS und InterCorp berechnen diesen Wert.

Für die übliche Nutzung der Korpora (wenn man nicht in die Tiefe der mathematischen Wahrscheinlichkeitsberechnung eintauchen will) zeigt sich bezüglich der Signifikanzmaße des Vorkommens sprachlicher Elemente der folgende Weg als leicht begehbar und hilfreich:
1) Für den Vergleich der Signifikanz unter verschiedenen Kollokationspartnern immer dieselbe Berechnung (z.B. LLR) verwenden. (Dies gilt als grundsätzliche Faustregel in der Forschung.)
2) Beim Untersuchen nur eines Kollokationspaars immer mehrere Berechnungen vergleichen.
3) Die Belege (Konkordanzzeilen mit den Kollokationspartnern) sollten immer stichprobenartig angeschaut werden, um zu überprüfen, ob die Ergebnisse der Berechnung nachvollziehbar sind.

Zu den weiteren wichtigen Eigenschaften der Korpora gehören **Annotationen**. Es sind zusätzliche Informationen zu den Korpustexten (**äußere Annotation**) und zu den einzelnen Wörtern im Korpus (**innere Annotation**).

Äußere Annotationen, genannt Metainformationen, beinhalten bibliographische Angaben zum Text und zu seinem Ursprung (Autor/in, Erscheinungsjahr, Quelle), aber auch andere, für die Untersuchung der Sprache wichtige Informationen wie Textsorte oder grobe stilistische Zuordnung. Für Gender-Studien beispielsweise sind bestimmt die Angaben über das Geschlecht des/der Verfassers/ in, Übersetzers/ in interessant. Die letztgenannten Angaben werden jedoch in den Korpora selten angegeben.

Innere Annotationen beziehen sich auf den Text: sie liefern Informationen zu jedem Wort im Text, über seine morphosyntaktischen Eigenschaften (**Tagging**) und syntaktisch-semantischen Rollen (**Parsing**).

Ein Tagger ist ein Programm, das durch **Tagging** jedem Wort im Text seine morphosyntaktische Eigenschaft zuweist: Der Tagger bestimmt die Grundform des Wortes, die Wortart, ggfs. auch andere morphosyntaktische Kategorien (Genus, Kasus, Numerus, Steigerungsform, Kategorie des Verbs, Tempus, Modus, in einigen Sprachen auch Aspekt). Das Ergebnis vom automatischen Tagging (ein Prozess, in dem der Text mit einem Tagger analysiert wird) ist in Abb. 4 und 5 zu sehen.

in */in/APPR* europäischer */europäisch/ADJA* Tradition */Tradition/NN* geführt */führen/VVPP* wird */werden/VAFIN* · /./$.
zur */zu/APPRART* Verantwortung */Verantwortung/NN* gezogen */ziehen/VVPP* werden */werden/VAINF* . /./$.

Abb. 4: Morphosyntaktisch annotierter und lemmatisierter Text - ein Auschnitt (InterCorp_de)

Parser ist ein Programm, das Sätze automatisch analysiert und einzelne Wörter im Satz syntaktisch annotiert, d.h. jedem Wort werden Informationen über seine syntaktisch-semantische Funktion im Satz zugewiesen (Abb. 5). Obwohl das Wort **Parsing** (Part-of-Speech-Analyse) treffender das Prinzip von Tagging charakterisiert, versteht man „unter Parsing […] heute eher solche Analyseprozesse, die substantiell über das bloße Annotieren eines Textes mit Wortarten hinausgehen und die grammatische Struktur einer Äußerung aufdecken“ (Carstensen et al. 2010: 303).

Parsing mit ParZu

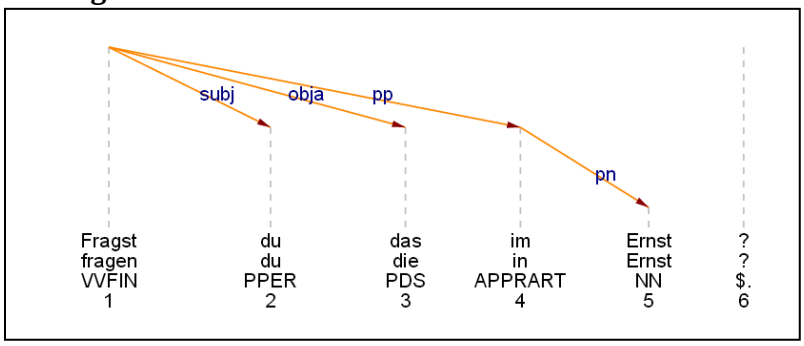

Abb. 5: Geparster Satz *Fragst du das im Ernst?* (ParZu)

Geparste Texte werden in **Treebanks** gespeichert. Treebanks (deutsch auch Baumbanken) sind Korpora, die das Suchen nach sprachlichen (semantischen und syntaktischen) Tiefenstrukturen ermöglichen. Die Texte müssen dafür vorher über einen Parser aufbereitet werden.

Neben den oben genannten Begriffen, die sich auf alle Korpora beziehen können, gibt es auch einige spezielle Fachbezeichnungen, die (ausschließlich) in Verbindung mit **Parallelkorpora** zu finden sind.

In **Parallelkorpora** werden übersetzte Texte in mehreren Sprachen eingespeist. Die Textsammlung für eine Sprache, zu der es in demselben Parallelkorpus Pendants noch in einer anderen Sprache gibt, heißt **Parallele**. Die sprachlichen Pendants/ Parallelen werden durch einen Zuordnungsprozess, sog. **Alignment,** einander zugewiesen.

Allgemein bedeutet **Alignment** die Anpassung und „Zuweisung“ zu einer anderen Parallele oder medialen Form des Textes. Im Parallelkorpus werden am häufigsten Text**segment**e auf der Höhe eines Satzes einander angepasst (Abb. 6). In multimedialen Korpora wird als Alignment die Synchronisierung des Textes mit dem Ton bezeichnet.

de	cs	en
Sie flogen kreuz und quer , um einander irrezuführen	Poletovali sem a tam , aby se navzájem svedli s cesty .	They flew hither and thither , to entice one another astray .
Herr Präsident ! In dieser Debatte geht es kreuz und quer	Pane předsedající , rozprava se pohybuje napříč různými směrnicemi .	Mr President , this debate criss-crosses the Directives .

Abb. 6: Segmente in drei Sprachparallelen (InterCorp_de)

Segmente sind Teile der Korpustexte, die nach unterschiedlichen Kriterien bestimmt werden – oft sind es Absätze oder Sätze. In einem Parallelkorpus handelt es sich um minimale Einheiten, die in zwei oder mehreren Sprachen aligniert (einander zugewiesen) werden, wie in der Abb. 6.

2.2 Wissenschaften rund um moderne Korpora

Die heutigen elektronischen Sprachkorpora könnte es nicht ohne Computer geben. Erst die Entwicklung der EDV hat es ermöglicht, große Datenmengen von sprachlichen Erscheinungen zu verwalten. Dazu entstand auch ein neues Fach an der Schnittstelle der Linguistik und Computertechnik: die Computerlinguistik.

Die **Computerlinguistik** „beschäftigt sich mit der maschinellen Verarbeitung natürlicher Sprache“ (Carstensen 2010: 1), sie versucht sprachliche Strukturen und Prozesse mit Hilfe der EDV-Technik zu beschreiben und zu simulieren (vgl. Metzler 2000: 129), sie unterstützt die Korpuslinguistik. Die Simulierung der menschlichen Sprache ist jedoch bisher mehr oder weniger gescheitert (z.B. Versuche der maschinellen Übersetzung). Die Beschreibung der Sprache erfolgt wiederum mit der Technik viel effektiver als früher. Man kann es mit einer Recherche in einem Bibliothekskatalog vergleichen: wie früher in Karteikatalogen der Bibliotheken gesucht wurde und wie heute mit elektronischen Katalogen recherchiert wird.

Die Anfänge der Computerlinguistik reichen bis in die ersten Jahre nach dem 2. Weltkrieg, in die USA, wo eben die ersten Versuche mit der maschinellen Übersetzung gemacht wurden. Der Begriff „computational linguistics“ wurde in den USA seit Anfang der 1960-er Jahre verwendet (Leech 1991: 8-10). Bis heute hat sich die Computerlinguistik als eine interdisziplinäre Wissenschaft (zwischen der angewandten Informatik und Linguistik) etabliert. Ihr Bereich reicht vom elektronischen Erfassen der Sprache über das Generieren von maschinellen Äußerungen, bis zur automatischen Transformation der geschriebenen zur gesprochenen (relativ einfach) Sprache und umgekehrt (bis heute nicht lückenlos gelöst).
Der wohl wichtigste Schwerpunkt der Computerlinguistik liegt in der **Korpuslinguistik**.

Die **Korpuslinguistik** ist jener Teil der Linguistik, der die Sprache systematisch anhand großer Mengen von realisierten sprachlichen Äußerungen (Parole) untersucht. Das Ziel des „korpuslinguistischen“ Ansatzes in der modernen Linguistik ist es, die Strukturen und Funktionen der natürlichen Sprache besser und realitätsnäher zu beschreiben (vgl. dazu Biber 1998).
Gerhard Budin (2011: 17) sagt zur Stellung der Korpuslinguistik in den Wissenschaften:

> „Die Corpuslinguistik ist eine interdisziplinäre, methodologische Grundlage für die computergestützte Forschung. Sie vernetzt im Kontext der Computerlinguistik die Sprachwissenschaft mit der Informatik und ermöglicht das Arbeiten auf einer breiten empirischen Basis.“

Dass diese „breite empirische Basis“ wichtig ist, lässt sich leicht am Beispiel aus einem renommierten Buch, das die Problematik der deutschen Wortbildung ausführlich behandelt, demonstrieren. In seiner Publikation *Wortbildung der deutschen Gegenwartssprache*“ schreibt Wolfgang Fleischer (1969: 117):

> „In Österreich sind z.B. geläufig die Formen (…) *Ausnahmszustand, Fabriksmarke, Gepäcksaufgabe, Zugsführer, Aufnahmsprüfung, Ausnahmszustand*; in der Schweiz z.B. (…) *Sportsmeldung, Zugsverbindung* u.ä.“

Eine einfache Recherche im Korpus bezeugt, wie frequentiert diese Formen im heutigen Deutsch sind: Im IDS-Korpus gibt es null Treffer/ Belege zum Wort *Ausnahmszustand*, dagegen zeigt die Streuung (das Vorkommen umgerechnet auf 1 Mio. Wörter) der Form *Ausnahmezustand* ganz deutlich, dass diese Form in allen drei Varietäten der deutschen Sprache gleich häufig verwendet wird. Genauso unüblich scheint das in der Schweiz „geläufige Fugen-S“, wie Fleischer (1969: 117) schreibt, im Wort *Sportsmeldung*: keine Treffer für *Sportsmeldung*, gegen 23 Treffer für *Sportmeldung*.

Die Korpuslinguistik kann auch die Theorien und Hypothesen der generativen Grammatik verifizieren oder widerlegen (vgl. Metzler 2000: 384). Dies wird aber von den meisten Generativisten abgelehnt (vgl. Lewandowski 1994: 612).

Die Korpuslinguistik ist keine neue „Wissenschaft", sondern nur ein Zugang, eine Sichtweise auf die Sprache. Zu Recht erntet sie Kritik seitens der Verfechter/-innen der mentalistischen, generativistischen Auffassung der Sprache (z.B. Chomsky 1986: 19), da kein Korpus alle Äußerungen einer Sprache erfassen kann. Auf der anderen Seite sieht man auch anhand von einigen wenigen Belegen, was in der Sprache überwiegt und typisch ist. Die Abstraktion des Typischen soll dann die Grundlage für die Grammatik der Sprache darstellen. Bisher wurden normative Grammatiken von nur wenigen Sprachen anhand von Korpusdaten erstellt. Selbst die letzte neubearbeitete Duden-Grammatik (2005) führt zwar Beispiele zu vielen grammatikalischen Erscheinungen aus dem Korpus DeReKo an, ignoriert jedoch sehr oft die Abweichungen in einzelnen Varietäten des Deutschen - so z.B. die Verbalkomplexe (2005: 481), wo die „österreichische" Reihenfolge trotz vieler Belege im Korpus (DeReKo) unerwähnt, dadurch implizit als falsch bezeichnet wird (dazu auch hier Studie 8). Viele offizielle Referenzwerke basieren auch heute noch auf ungenügenden Belegsammlungen bzw. immer noch auf dem „Gefühl" ihrer Autor/-innen (nicht anders ist es im Fall Österreichisches Wörterbuch, teilweise auch Duden Deutsches Universalwörterbuch u.a.) Es muss betont werden, dass es im korpuslinguistischen Ansatz auf keinen Fall darum geht, was in der Sprache korrekt oder inkorrekt ist, sondern darum, welches Phänomen in der Verwendung der Sprache typisch ist, welches untypisch, selten oder rar ist. In diesem Zusammenhang darf der Korpuseinsatz in DaF/DaZ-Unterricht und -forschung nicht unerwähnt bleiben. DaF/DaZ passiert eben auch nicht in einer „idealen Sprachwelt" – es ist jeweils in einem konkreten Kontext eingebettet, der bei der Vermittlung des Deutschen berücksichtigt werden muss. Erst in diesem Kontext sucht man das Typische, bzw. das für die Lerner/-innen Hilfreiche. Dem soll auch die Wahl des passenden Instruments entsprechen. So gesehen grenzt die Korpuslinguistik auch an die Methodologie der Fremdsprachenvermittlung – diese Schnittstelle wurde allerdings noch wenig beschrieben und erforscht.

Die Vorreiter des korpusunterstützten Fremdsprachenunterrichts waren Artikel, Bücher, sogar kleine Wörterbücher über Fehler, die die Lerner/-innen wiederholt begehen, z.B. Swans´ (1982 u. weitere Auflagen) *Practical English usage*; Wörterbücher der Falschen Freunde bei deutschen Englischlernenden (z.B. von Barnickel (1992)); Sparlings´ (1989): *English or Czenglish* über die Interferenzfehler im Englischen bei tschechischen Englischlernenden u.a.m. Mit den EDV-Möglichkeiten kann sich eigentlich jede/-r ein Korpus erstellen, in dem effektiv gesucht werden kann. Ein aufwendigeres Instrument für die Fehlerforschung (Falko) ist bereits an der Humboldt-Universität zu Berlin entstanden (siehe Reznicek et al. 2012). Wie man sich selbstständig ein Korpus einfach erstellen kann, wird im Kap. 5.1 (TextSTAT) beschrieben.

An und für sich ist die Erstellung eines Korpus und die Arbeit mit ihm nichts Neues (Francis 1992: 17). Die ersten Korpora entstanden als Basis für die Zusammenstellung von Wörterbüchern und Grammatiken – in Form von unzähligen Karteizetteln. Zum Beispiel die traditionelle Sprachkartei der Dudenredaktion verfügte über drei Millionen authentischer Sprachbelege (Drosdowski 1985: 87). Die Anfänge der Korpuslinguistik reichen also tief in die Geschichte der Sprachforschung.

Die heutige Korpuslinguistik arbeitet ausschließlich mit elektronischen Sprachkorpora und zeigt so den Trend in der modernen Philologie an. Ihre Bedeutung ist vor allem im

Zusammenhang mit dem Aufkommen der EDV- Technik angestiegen. Mit der Verbreitung des Internets ist diese moderne Methode der Sprachwissenschaft und der Sprachvermittlung für beinahe jedermann zugänglich.

3. Korpustypologie

Die Vielfalt der Korpora lässt sich etwa mit der Vielfalt der Wörterbücher und Enzyklopädien vergleichen: für jeden Zweck gibt es das passende Werk, man muss nur wissen, was man braucht. Wer die Bedeutung eines Wortes sucht, greift zum Bedeutungswörterbuch, äquivalente Bezeichnungen in einer anderen Sprache findet man in einem Übersetzungswörterbuch. Wer etwa mehr über eine Sache wissen will, nimmt eine Enzyklopädie zur Hand. Heutzutage wird die Suche noch erleichtert und beschleunigt, denn die meisten modernen Lexika gibt es auch in elektronischer Form.

Ähnlich ist es mit elektronischen Korpora: Wer im Internet sucht, der wird oft fündig. Und wenn kein passendes Korpus existiert, lässt es sich irgendwie erstellen. Eine selbstständige Erstellung von einsprachigen Korpora geht relativ leicht z.B. mit dem Programm TextSTAT (siehe Kap. 5.1). Das Erstellen eines Parallelkorpus ist schon aufwendiger, aber auch nicht unmöglich (z.B. mit dem kostenpflichtigen Programm ParaConc von M. Barlow (2009)). Auf jeden Fall müsste man bei diesen in Selbsthilfe erstellten Korpora auf viele hilfreiche und angenehme Eigenschaften und Tools verzichten. Deshalb sehe ich die Nutzung der bereits bestehenden, oft professionell erstellten Korpora als den sinnvollsten Weg.

Eine übersichtliche Graphik mit Kriterien, nach denen Korpora (allgemein) aufgeteilt werden können, präsentieren Lemnitzer/ Zinsmeister (2010: 103). Eine ähnliches Schema (ergänzt um einige Eigenschaften, auf die man bei der Korpuswahl Rücksicht nehmen soll) präsentiert auf seiner Homepage auch Stephen Berman (2013).

Die folgende Typologie reißt die Grundaufteilung der Korpuslandschaft auf, dabei werden auch zu jedem Typ die Namen der entsprechenden Korpora aufgelistet. Die Vertreter der einzelnen Korpustypen werden hier in Abkürzung angeführt. Eine Übersicht mit vollen Namen und Internetadressen befindet sich ab der Seite 219.

Alle hier erwähnten Korpora sind einfach zugänglich. Sie können m.E. den DaF/DaZ-Bereich in der Lehre und Forschung unterstützen.

Die einzelnen Korpustypen schließen einander nicht aus. Viele Korpora können verschiedenen Zwecken dienen (diese werden bei einigen Korpora im Kap. 4 näher charakterisiert) - je nachdem, wie das Korpus aufgebaut ist, welche Eigenschaften der Korpusmanager hat und welche Abfrage möglich ist.

3.1 Stadium der gespeicherten Sprache

3.1.1 Synchrone Korpora

Korpora der Gegenwartssprache (synchrone Korpora) beinhalten aktuelle Texte der letzten Dekaden. Meist sind es Texte, die nach dem 2. Weltkrieg entstanden sind. Als Schwelle zur „heutigen“ Sprache gilt das Jahr 1989 (Kocek et al 2000: 13).

Einen nicht geringen Teil der nationalen Korpora bilden aus pragmatischen Gründen auch Texte vor 1945, v.a. belletristische Texte, wegen ihrer Sprache, die (immer noch) als modern gilt (Kafka, Böll, sogar Hašek und sein Švejk). Im Unterschied zu Gebrauchstexten veraltet

die Belletristik wesentlich langsamer – u.a. auch dank der überarbeiteten neuen Auflagen der alten Werke.

Synchrone Korpora mit Deutsch (nach Textvolumen absteigend)**:**

geschriebene Sprache: **DeReKo**, **deTenTen**, **Wortschatz Leipzig**, **DWDS**, **deWac**, **InterCorp_de**, **OPUS;**

gesprochene Sprache: **DGD, DWDS – gesprochene Sprache, GeWiss, BAS**

3.1.2 Diachrone Korpora

Mit Diachronie bezeichnete de Saussure einzelne aufeinander folgende Entwicklungsstadien der Sprache. Diachrone Korpora sollen helfen, die Sprache(n) auf der Zeitachse zu betrachten, Änderungen und Tendenzen in der Entwicklung der Sprache/den Sprachen besser zu verstehen. Jedes Korpus, das Texte aus mehreren Jahrzehnten beinhaltet, und dessen Tools es ermöglichen, die Texte nach der Zeit ihres Entstehens/ Herausgebens zu filtern, kann als ein diachrones Korpus bezeichnet werden.

Diachrone Korpora des Deutschen (letzte Jahrhunderte):

DWDS (20. Jh.), **DeReKo** (18. – 21. Jh.)

3.1.3 Historische Korpora

In die historischen Korpora werden Texte aus einem älteren, bereits überwundenen Stadium der Sprache eingespeist. Sie dienen in erster Linie als Forschungsquellen für das Studium der jeweiligen Entwicklungsphase der Sprache. Darüber hinaus fungieren sie als eine Datenbank von historischen Texten, die sonst nur schwer zugänglich wären. Für DaF (jedoch eher im universitären Bereich, speziell der Auslandsgermanistik) haben sie insofern Bedeutung, da zum vollkommenen Studium der Sprache auch ein Einblick in ihre historische Entwicklung gehört. Für DaZ spielen sie keine Rolle.

Historische Korpora des Deutschen (nach Entwicklungsstadium der Sprache)**:**

Althochdeutsch: **DDD** (Deutsch Diachron Digital: Referenzkorpus Altdeutsch)

Mittelhochdeutsch: **MHDBDB** (Mittelhochdeutsche Begriffsdatenbank)

Frühneuhochdeutsch: **FnhdC** (Das Bonner Frühneuhochdeutschkorpus)

alle historische Stadien: **DDD** (Deutsch Diachron Digital)

Die Begriffe „diachrone“ und „historische“ Korpora werden oft willkürlich und synonym verwendet. Die richtige Bezeichnung für historische Korpora ist „Korpora mit historischen Texten“. Alle Korpora, in denen Texte aus mehreren Jahren oder Dekaden gespeichert sind, können eigentlich als „diachrone Korpora“ bezeichnet werden.

3.2 Medium

3.2.1 Korpora der geschriebenen Sprache

Die geschriebene Sprache lässt sich viel einfacher digitalisieren und in eine (von einem Korpusmanager) lesbare Form umwandeln als die gesprochene Sprache. Deswegen haben Korpora der geschriebenen Sprache viel mehr Vertreter in der Korpuswelt. Korpora der geschriebenen Sprache beinhalten digitalisierte authentische Texte, die in irgendeiner geschriebenen Form (gedruckt oder elektronisch) öffentlich erschienen sind.

Elektronisch verfasste Texte bilden das Gros aller großen Korpora, viele Texte werden jedoch immer noch eingescannt.

Korpora der geschriebenen deutschen Sprache:

DeReKo, **deTenTen** (nur Internettexte), **Wortschatz Leipzig** (nur Internettexte), **DWDS**, **deWac** (nur Internettexte), **InterCorp_de**, **OPUS** (nur Internettexte)

3.2.2 Korpora der gesprochenen Sprache

Obwohl (immer noch) mehr gesprochen als geschrieben wird, bilden Korpora der gesprochenen Sprache nur einen Bruchteil in der Korpuslandschaft aller Sprachen. Für viele Sprachen gibt es sie noch gar nicht. Der Grund liegt in ihrer aufwendigen technischen Aufbereitung: Korpora der gesprochenen Sprache sind transkribierte und digitalisierte gesprochene Texte. Das weitere Manko dieser Instrumente ist ihre Authentizität (ebenso wie häufig auch bei Hörtexten in Lehrbüchern): die Sprechereignisse sind oft gesteuert, künstlich aufgebaut (teilweise **DGD**), oder repräsentieren gar das Vorlesen eines vorgefertigten, geschriebenen Textes (**adaba**). Für die Erforschung der Aussprache sind aber auch diese Vorgangsweisen legitim.

Das Erstellen eines repräsentativen Korpus der gesprochenen Sprache scheitert grundsätzlich an zwei Punkten: 1) die heutige Technik ist noch nicht so weit, alle Nuancen in der Aussprache unterschiedlicher Sprecher zu erfassen (v.a. wenn es sich auch um dialektale Färbungen handelt); 2) das Aufnehmen von Privatgesprächen betrachtet die europäische Gesetzeslage als widerrechtlich, falls es ohne ausdrückliches Erlaubnis der beteiligten Personen erfolgt[10].

Korpora der deutschen gesprochenen Sprache:

DGD, **BAS, DWDS - gesprochene Sprache**

Lemnitzer/ Zinsmeister (2010: 103) listen noch „multimodale Korpora“ auf – darunter verstehen sie Korpora mit Audio- und/oder Videoaufnahmen. Nach dieser Definition sind alle hier angeführten Korpora der gesprochenen Sprache „multimodal“, da jeweils auch Audiofiles abrufbar sind.

3.3. Repräsentativität

3.3.1 Stilistische Kriterien

Allgemeine Korpora

Allgemeine Korpora versuchen Texte aller Textsorten aus (möglichst) vielen Kommunikationsbereichen und stilistischen Ebenen abzudecken.

Allgemeine Korpora mit Deutsch:

DeReKo, **DWDS**, **deTenTen**

Spezifische Korpora

Spezifische Korpora werden für die Erforschung nur eines kleineren Segments der Sprache aufgebaut. Sie beinhalten Texte, die ein gemeinsames Spezifikum verbindet - z.B. Fachbereich (medizinische oder naturwissenschaftliche Texte …), Stil (journalistische oder

[10] Meine Redakteurin meint, ich soll es begrüßen, was ich grundsätzlich auch tue. Dennoch schadet dieses Gesetz der Erstellung der Korpora der gesprochenen Sprache.

dramatische Texte, Texte im Dialekt…), Autor (Goethe, Orwell…), andere besondere Zwecke (Fehlerkorpus, Lernerkorpus, Varietätenkorpus…)

Spezifische Korpora mit Deutsch:

Viele größere Korpora ermöglichen auch ein spezifisches Korpus als Subkorpus zu erstellen und nur in diesem zu recherchieren. Beispielsweise im **DeReKo** gibt es das Korpus **GOE** mit Goethes Werken oder ein (allerdings nicht öffentliches) Korpus **IKO**, wo ausschließlich Zeitungsinterviews gespeichert sind u.v.a.m. (siehe Cosmas II → Hilfeportal → Organisation des Textmaterials).

Deutsch ist vertreten auch in verschiedenen spezifischen Parallelkorpora – z.B. im MULTEXT-East corpus (George Orwells Novelle 1984 in 11 Sprachen), im OPUS-Corpus (Filmuntertitel, EU-Verfassung).

Weitere spezifische Korpora:

Varietätenkorpus: **Korpus C4** (im Aufbau)

Fehlerkorpus (Lernerkorpus): **Falko**

3.3.2 Kriterium der Ausgewogenheit

Ausgewogene Korpora

Über genaue Kriterien für die Erstellung eines ausgewogenen Korpus herrscht kein Konsens. Beispielsweise folgen zwei repräsentative „Nationalkorpora" unterschiedlichen Ausgewogenheitskriterien: Im DeReKo überwiegen Zeitungstexte, im ČNK (Referenzkorpus SYN 2005) jedoch Belletristik.

In einem ausgewogenen, repräsentativen und allgemeinen Korpus soll das Verhältnis der Textsorten und Textstile dem Verhältnis in der realen sprachlichen Welt entsprechen. Das heißt, dass den größten Teil der Korpora unbedingt die gesprochene Sprache bilden müsste. Da dies (noch) nicht möglich ist (siehe oben), werden als ausgewogene Korpora solche bezeichnet, die ein stabiles Verhältnis der einzelnen geschriebenen Textsorten beinhalten. Große nationale Korpora werden laufend um Textsorten ergänzt um ihre Ausgewogenheit zu gewährleisten.

Deutsche Korpora mit repräsentativ ausgewogenen Daten:

DeReKo, **DWDS**

Opportunistisch gebildete Korpora

Es sind Korpora, die keinen Anspruch auf Abdeckung der Kriterien der Ausgewogenheit haben, bzw. sie erfüllen diese Kriterien nicht.

Nicht ausgewogene Korpora mit Deutsch

ausschließlich Internettexte: **Wortschatz-Portal**, **deTenTen**, **deWac**

Protokolle, Festschriften: **OPUS-Corpus**

Belletristik und Publizistik: **InterCorp**

3.4 Größe

Das Kriterium der Korpusgröße hängt unmittelbar mit der Repräsentativität der gespeicherten Daten und dadurch auch mit der Validität der Ergebnisse, die man aus Abfragen gewinnt, zusammen. Grundsätzlich gilt, dass kleinere oder kleine Korpora[11] zwar übersichtlicher, daher auch (oft) fehlerfrei sind, sie aber für die Erforschung oder Darstellung allgemeiner

[11] Ein kleines Korpus ist z.B. **DeuCze** (siehe Kotulková et al. 2005-2010).

sprachlicher Phänomene keine repräsentativen Daten liefern können. Gut können sie nur ein kleines, spezifisches Segment oder ein hoch frequentiertes Element der Sprache aufdecken.

Ein kleines Korpus lässt sich grundsätzlich aus jedem größeren Korpus (als Sub-Korpus) erstellen. Im korpuslinguistischen Ansatz der Sprach(en)forschung und -vermittlung gilt: je größer das Korpus, desto repräsentativer, sicherer, überzeugender sind die Rechercheergebnisse. Die Grundstrukturen und die hochfrequentierten Elemente der Sprache lassen sich zwar tatsächlich schon anhand von kürzeren Texten erfassen, wie aus dem Kap. 5.1 ersichtlich ist, viele Fragen (z.B. über die Lexik, Peripherie der Sprache, Dialekte) können aber nicht einmal die jetzigen Korpora mit Milliarden Wörtern beantworten.

Sehr große Korpora mit Deutsch:

DeReKo (8,9 Mrd. Wörter), **deTenTen** (1 Mrd. Wörter), **deWac** (1,4 Mrd. Wörter)

3.5 Sprache

3.5.1 Monolinguale Korpora

In einsprachige (monolinguale) Korpora werden Texte einer Sprache eingespeist, ungeachtet dessen, ob es sich um Originaltexte oder um Übersetzungen aus anderen Sprachen handelt. Korpora mit plurizentrischen Sprachen (Deutsch, Englisch, Spanisch) haben entweder überhaupt selbstständige Instrumente für die jeweilige Varietät (DWDS, Schweizer Textkorpus) oder - falls sie „unter einem Dach“ verwaltet werden - manchmal Tools, die das Filtern von einzelnen Varietäten ermöglichen (DeReKo). Viele Korpora ermöglichen dies jedoch nicht (Wortschatz-Portal, deWac).

Monolinguale Korpora mit Deutsch:

DeReKo (D-A-CH), **AAC** (nur Österreich), **DWDS** (nur Deutschland), **Schweizer Textkorpus** (nur Schweiz)**; DGD**, **deWac**, **deTenTen**, **Wortschatz-Leipzig**.

Nationalkorpora anderer Sprachen siehe Kap. 4.1.

3.5.2 Bilinguale und Multilinguale Korpora

Zwei- und mehrsprachige Korpora können theoretisch noch in zwei Kategorien unterteilt werden: **Vergleichskorpora** und **Parallelkorpora**.

In **Vergleichskorpora** findet man Texte zum selben Thema, wobei diese Texte unabhängig voneinander als Originale entstanden sind. Das heißt, zwischen zwei Texten des Vergleichskorpus wurde keine übersetzerische Arbeit geleistet. Als Vergleichskorpus könnten beispielsweise Nachrichten zum selben Thema in unterschiedlichen Medien und Sprachen dienen. Als große Quelle von Texten für ein Vergleichskorpus können auch diejenigen Artikel in Wikipedia dienen, die nicht als Übersetzungen gekennzeichnet sind (vgl. Wikipedia: Übersetzungen (2014)).

Parallelkorpora sind Datenbanken mit „offiziellen“ Übersetzungen, also mit denselben Texten, die in mehreren Sprachen existieren und ihre kommunikative Funktion erfüllen.

Beim Vergleich dieser beiden Subkategorien eröffnet sich automatisch die Frage nach der (1) **Vergleichbarkeit der Texte**: können wir zwei unabhängig voneinander verfasste Texte als zwei gleiche sprachliche Handlungen betrachten? Können wir auch Übersetzungen aus „dritten Sprachen“ vergleichen (z.B. deutsche und englische Übersetzungen des ungarischen

Schriftstellers István Örkény „One Minute Stories/ Minutengeschichten“); (2) **Authentizität der Texte**: eine Übersetzung ist kein frei formulierter Text. Darf er also die „reale“ Sprache repräsentieren?; (3) **Kompetenz der Übersetzer/-innen** und die damit verbundene Gefahr der Übersetzungsfehler.

Diese Fragen muss man bei der Arbeit mit Parallelkorpora immer im Auge behalten. Es sind eben Faktoren, die das Bild (d.h. die Ergebnisse der Recherche) verzerren können. Auf der anderen Seite gibt es wohl keine andere Möglichkeit, Sprachen untereinander zu vergleichen. Wir können uns nur darauf verlassen, dass die Menge und Vielfalt der Texte im Korpus die genannten Fehlerquellen wieder ausgleichen.

Bilinguale und multilinguale Korpora mit Deutsch:

InterCorp, OPUS-Corpus

3.6 Technische Eigenschaften

Eigentlich sind es die technischen Eigenschaften, die aus einer simplen Textsammlung ein Korpus machen. In diesem Punkt unterscheidet sich auch jede Korpusarbeit vom Googeln.

Viele Korpustools und Zusatzinformationen sind auf den ersten Blick irrelevant. Sie sind aber für die Suche im Korpus sehr wichtig. Darüber hinaus helfen sie besser zu verstehen, wie die Korpora funktionieren und aufgebaut sind. Nach Berman (2013, ergänzt) lassen sich folgende korpustechnische Eigenschaften nennen:

- **Metadaten**: Informationen über den Aufbau und Inhalt des Korpus. Nach einzelnen Posten der Metadaten lassen sich Subkorpora erstellen: z.B. Texte aus einer Zeitperiode (DeReKo, DWDS), eines Autors (InterCorp), Texte, die nur Frauen übersetzt haben (InterCorp) u.a.m.

- **Linguistische Aufbereitung**: hinzugefügte linguistisch relevante Markierungen im Korpus
    - keine: rohe sprachliche Daten (Texte)
    - **Tokenisierung:** Segmentierung der Texte in kleinere Einheiten: Absätze, Sätze, Wörter und Satzzeichen (alle gängigen Korpora); Phrasen, Konstituenten (geparste Korpora)
    - **Lemmatisierung:** Zuweisung der flektierten Formen zu einer Grundform (lemmatisierte Korpora)
    - **Annotation**: Markierung von (morpho)syntaktischen, semantischen u.a. Eigenschaften:
        - **Tagging** (Morphosyntax, z.T. auch Semantik): Wortart, Flexionsmerkmale, Eigennamen (getaggte Korpora)
        - **Parsing** (Syntax): Konstituenten, syntaktische Funktionen, topologische Felder (geparste Korpora).
        - **Phonetik/Prosodie/Intonation** (DGD)
        - **Fehler**: evtl. auf allen Ebenen, z.B. Rechtschreibung, Kongruenz, Wortstellung, Tempus usw., dienlich für Sprachunterrichts- sowie Spracherwerbsforschung (Falko)

- **Word Sketches:** automatische Berechnung und Darstellung der Eigenschaften und des Verhaltens eines Wortes im Korpus. (deTenTen, teilweise auch DWDS)

Zu bemerken ist, dass all diese technischen Korpusfinessen mehrheitlich automatisch durchgeführt werden, dadurch stößt man immer wieder auf fehlerhafte Angaben. Dies ist ein Preis für die große Menge der Daten, die mit manueller Aufbereitung gar nicht oder nur mit einem enormen Aufwand freigegeben werden könnten.

4. Korpora relevant für DaF/DaZ

Der folgenden Auswahl mit einer kurzen Beschreibung der einzelnen Korpora liegen diese Überlegungen zu Grunde:

1) Brauchen DaF/DaZ-Lehrer/-innen und Schüler/-innen überhaupt Korpora? Wenn ja, wozu?
2) Von welchen Korpora könnten sie am meisten Gebrauch machen?
3) Welche Korpustools und Korpuseigenschaften sind am ehesten einsetzbar und hilfreich im DaF/DaZ-Unterricht?

Gleich die erste Frage ist schwierig zu beantworten: Für die Sprachenvermittlung ist an sich natürlich kein Korpus notwendig. Der Sprach(en)unterricht läuft in den meisten Fällen auch heute ohne Korpora und mit Erfolg. Es wäre auch nicht sinnvoll Korpora im Unterricht zwanghaft einzusetzen. Die Arbeit mit fast allen Korpusmanagern (also die Bedienung dieser Instrumente) erfordert einige technische und linguistische Vorkenntnisse. Die Interpretation der gewonnenen Daten benötigt auch gewisse Erfahrungen mit dem Sprachgebrauch. Dies alles darf man von vielen Kursteilnehmer/-innen (vor allem im DaZ-Bereich) nicht erwarten. Darüber hinaus ist auch das sichere Erlernen der Korpusbedienung zeitaufwändig, noch dazu werden die meisten Korpusmanager verbessert, das Bild der Schnittstelle verändert sich und so könnte man leicht verzweifeln und skeptisch werden, weil man etwas gelernt hat, plötzlich aber alles anders ist und man wieder wie am Anfang sucht.
Auf alle Fälle ist es aber dennoch wichtig, zumindest die Information über die Existenz von Korpora zu vermitteln und diese als eine zuverlässige Informationsquelle über die Sprache darzustellen.

Korpora können und sollen den (Fremd-)Sprachenunterricht nicht steuern, sie können und müssen ihn unterstützen, denn die kommunikativ-pragmatische Methode des modernen Sprachunterrichts erfordert die Arbeit mit authentischen Texten, die Vermittlung realer Sprachmittel und häufiger Strukturen. Das alles ist in einem natürlichen Kontext im Korpus explizit oder implizit beinhaltet.

Die meisten DaF-Schüler/-innen sind im Umgang mit Computern gewandt. Bei den DaZ-Lernenden (v.a. bei den Erwachsenen) ist dies nicht automatisch zu erwarten. Für sie könnte die Arbeit mit einem elektronischen Instrument zwar eine (zusätzliche) Belastung sein. Auf der anderen Seite könnte es für sie auch eine positive Herausforderung darstellen, sich mit ihrer eigenen Sprache zu befassen und diese vielleicht mit dem Deutschen zu vergleichen – vorausgesetzt, dass es dazu entsprechende Korpora gibt. Beim jetzigen Stand der DaZ-Klientel klingt dieses Statement etwas naiv, man soll jedoch bedenken, dass sich das Korpusangebot stets erweitert.

Selbst die Grundkenntnisse über Korpora und die Fähigkeit, die einfachsten Recherchen durchführen zu können, unterstützen das autonome Lernen sowohl im induktiven als auch im deduktiven Weg: Lernende können selbst Regelmäßigkeiten in der Sprache aufdecken, sie können auch bereits gelernte Strukturen überprüfen. Dabei sollen sie sich aber bewusst sein, dass sie sich in der realen Sprache bewegen, in der die Theorie immer wieder an ihre Grenzen stößt. Diese Tatsache soll auch zur kritischen Stellungnahme zu theoretischen Grundlagen und Darstellungen in Grammatiken führen.

Bei der Wahl aus dem Korpusangebot müssen wir uns damit begnügen, was es gibt. Es gibt schon viele nutzbare Korpora, allerdings klaffen in der Landschaft auch ziemlich viele Löcher, die noch gefüllt werden müssen.

Falls wir Korpora im DaF/DaZ-Unterricht einsetzen wollen, wäre es sicher sinnvoll auch die nationalen (einsprachigen) Korpora der jeweiligen Ausgangssprache zu erwähnen. Auch im DaZ-Unterricht wäre es sinnvoll, sich einen Überblick zu verschaffen, ob es vielleicht nicht ein Korpus für die Erstsprache der Lernenden gibt.

In europäischen Dimensionen sollen wir zumindest eine grobe Vorstellung davon haben, welche Korpora unsere Nachbarsprachen zur Verfügung haben.

Die eingesetzten Korpora und Korpusinstrumente sollen zugänglich, möglichst repräsentativ und relativ einfach und übersichtlich sein. Die Bedienung vieler Korpora sieht auf den ersten Blick zwar ziemlich komplex aus und kann von der Arbeit fast abschrecken (DeReKo, InterCorp), gut gewählte Beispiele von einfachen Abfragen können jedoch die Arbeit vorentlasten und so bildhaft darstellen, wie schnell man eigentlich Vieles über den Sprachgebrauch erfahren kann – wie das folgende Beispiel verdeutlicht. Die Abfrage (Korpus InterCorp_de) lautete: **ich rede.** Aus mehreren Dutzend Konkordanzen werden hier nur einige präsentiert, um die Übersichtlichkeit zu bewahren:

(1)	» Kannst du bitte die Karte weglegen , während	**ich rede**	? « » Entschuldigung . «
(2)	ich habe eine Frau , aber	**ich rede**	kaum mit ihr ,
(3)	(…) , und ich weiß , wovon	**ich rede**	.
(4)	Aber	**ich rede**	jetzt von dir .
(5)	Mein Gott ,	**ich rede**	ja auch manchmal solche Blödheiten , …
(6)	» Ich weiß , mit wem	**ich rede**	! «
(7)	Du musst ja wissen , wovon	**ich rede**	.
(8)	… aber je länger	**ich rede**	, desto ärgerlicher werde ich .
(9)	Ja ,	**ich rede**	jetzt schon zu lange, ich weiß …
(10)	Ich möchte Ihnen mitteilen - und	**ich rede**	hier über Fakten , nicht Träumereien … -

Abb. 7: Konkordanzzeilen zur Abfrage: Phrase *ich rede* (InterCorp_de)

Auch ein/e Anfänger/-in kann sich ein Bild machen, wie die Verbindung *ich rede* in unterschiedliche Sätze eingebettet werden kann: *ich rede* (1); *ich rede* (*etwas*) (5); *ich rede* (*wie*) (8), (9); *ich rede mit* (*jemandem*) (2), (6); *ich rede von* (3), (4), (7); *ich rede über* (10).

Diese Überlegungen zur Sinnhaftigkeit des Korpuseinsatzes münden in die folgende Auflistung einiger (nationaler) Korpora von Sprachen, die in Mitteleuropa am häufigsten anzutreffen sind. Gemeint sind Korpora der offiziellen Landessprachen einzelner mitteleuropäischer Länder (Französisch, Polnisch, Slowakisch, Tschechisch, Ungarisch), aber auch Sprachen, die viele Migranten/-innen sprechen (Russisch, Griechisch, Türkisch). Selbstverständlich werden auch einige Korpora des Englischen erwähnt. Darauf folgt eine detailliertere Übersicht über Korpora, in denen die deutsche Sprache gespeichert ist.

4.1 NATIONALE KORPORA (außer mit Deutsch)

Die (nationalen) Korpora der einzelnen Sprachen sind hier alphabetisch aufgelistet. In die Auswahl kamen Korpora der Nachbarsprachen der deutschsprachigen Länder und anderer wichtiger Sprachen. Ihre Namen sind zuerst linear (wortwörtlich) ins Deutsche übersetzt, dann folgen ihre offiziellen Namen in der Landesprache und – falls vorhanden – auch die offizielle englische Bezeichnung. Vertreten sind „große“, aber auch „kleinere“ Sprachen. Die Auswahl der Korpora unterliegt folgenden Kriterien: Zugänglichkeit, Aufwendigkeit. Hier werden nur leicht zugängliche Korpora erwähnt, in denen einfach und direkt (über eine Internetschnittstelle) recherchiert werden kann. Nach dem Bild der Startseite[12] wird als Beispiel für die Arbeit mit dem jeweiligen Korpus ein Bild mit einigen Konkordanzzeilen angeführt. Die Konkordanzen werden durch die entsprechende **Abfrage** eingeleitet: in Klammer stehen Buttons oder Suchfeldnamen (in der Landessprache oder auf Englisch), in die die Abfrage eingegeben wird, bzw. die Beschreibung des Weges zu ihnen. Nach dem Doppelpunkt steht die Abfrage in der Abfragesyntax, die der Korpusmanager verlangt.

4.1.1 Englisch

Britisches Nationalkorpus (BNC)/ British National Corpus

http://www.natcorp.ox.ac.uk/

Das Britische Nationalkorpus beinhaltet 100 Mio. Wörter: 90% geschriebenes und 10% gesprochenes Britisches Englisch „aus den späten Jahren des 20. Jahrhunderts“ (BNC 2009 → About BNC).

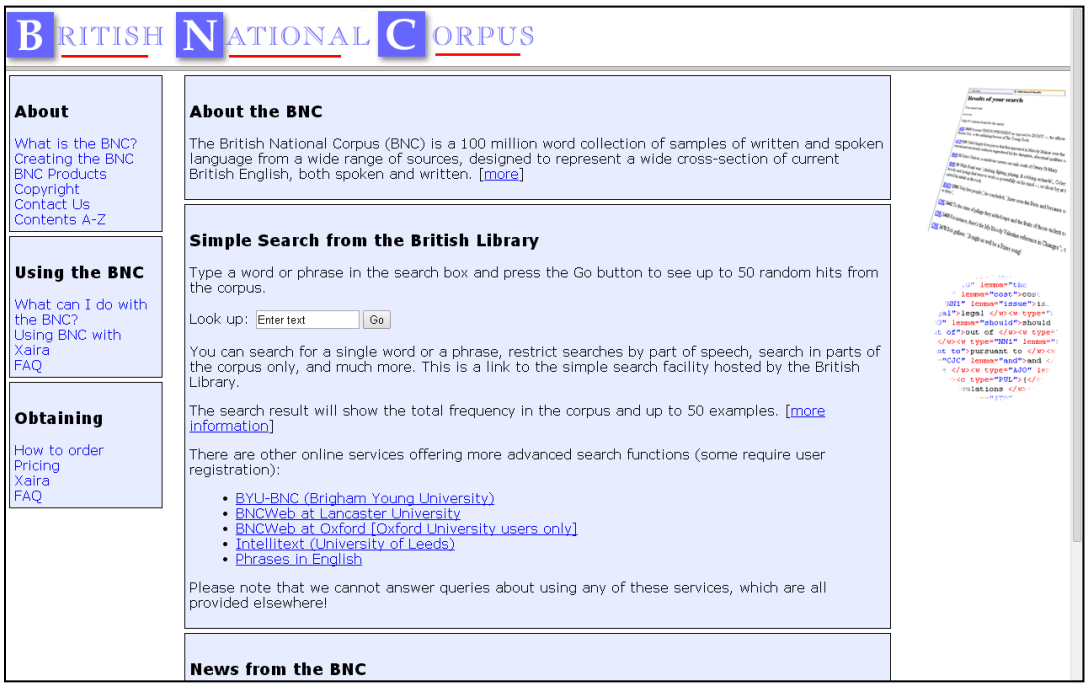

Abb. 8: BNC-Startseite (2014)

[12] Stand Frühjahr 2014 - dieses wird sich mit großer Wahrscheinlichkeit mit der Zeit ändern, da auch die Korpusseiten einem regelmäßigen Lifting der graphischen Gestaltung unterzogen werden.

Abfrage
Look up: **school**

Results of your search

Your query was
school

Here is a random selection of 50 solutions from the 37146 found.

AL8 **114** In 1926, eleven of her students attended Newnham College's summer school for working women and a further three joined the summer **school** at Cheshunt College, where George Pateman and his wife were in charge of the arrangements.

AM7 **417** An example of this would be the operation of a lettings policy in a **school**.

AMB **1233** The library was one of the most confusing places in the **school**.

Abb. 9: Konkordanzen zur Abfrage Wortform *school* (*Schule*) (BNC)

Volles Korpusservice (Suchmöglichkeiten nach unterschiedlichen Attributen, statistische Funktionen etc.) ist mittels Xaira-Programm möglich. Download und Manual dazu sind auf der Homepage des BNC zu finden. Die Daten des BNC sind auch (teilweise) unter InterCorp recherchierbar.
Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

Andere Korpora mit Englisch

Andere Korpora mit Englisch sind auf mehreren Web-Seiten aufgelistet. Siehe z.B. unter Braun (2005) oder Davies (2014).

Ein interessantes Korpus ist bestimmt das **GloWbE (Corpus of Global Web-based English:** http://corpus2.byu.edu/glowbe/ **)** mit fast 2 Mrd. laufenden Wörtern aus 20 Ländern, in denen Englisch als eine übliche Kommunikationssprache verwendet wird. Der Korpusmanager erlaubt Abfragen, bzw. eine Darstellung der Ergebnisse nach einzelnen Ländern, daher ist es auch ein wertvolles Instrument für die Untersuchung der einzelnen Varietäten des Englischen. Im Korpus sind ausschließlich Texte aus Internetseiten gespeichert.

Englisch ist auch in vielen Parallelkorpora vertreten.

4.1.2 Französisch

Textkorpus FRANTEXT/ Base textuelle FRANTEXT

http://www.frantext.fr/ctlf/
Die „Textdatabase“ (wörtliche Übersetzung) beinhaltet Texte geschriebener französischer Sprache vom 12. bis zum 21. Jahrhundert mit dem Umfang von über 271 Mio. Wörtern. Die Schnittstelle sowie alle Informationen (bis auf eine kurze Projektdarstellung in Englisch) sind ausschließlich auf Französisch.

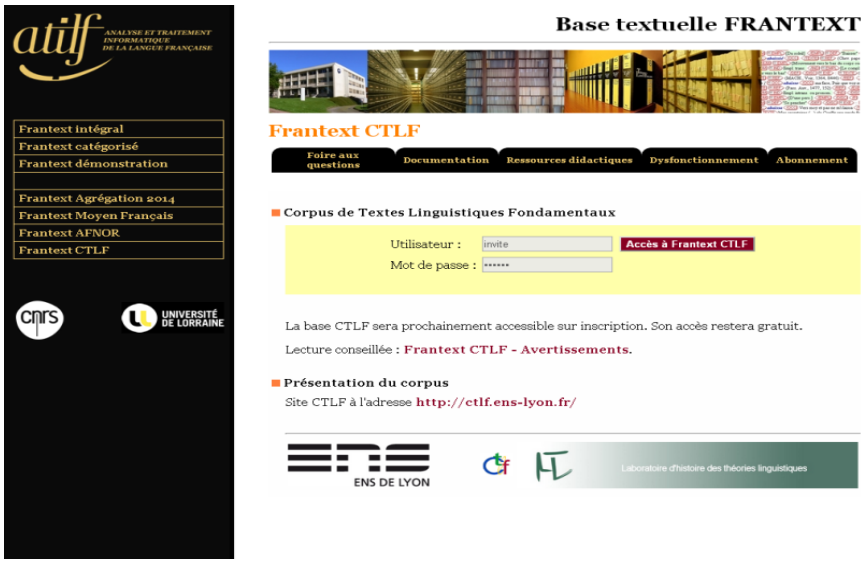

Abb. 10: Frantext-Startseit (2014)

Abfrage
Mot à rechercher: **école** expression régulière

Résultats 1 à 20 / 557				
[1]	C002	maniere, laquelle fait reconnoître la main ou l'	**école**	de. l'artiste. L'observation montrant de
[2]	C002	compositions : raison pourquoi ils avoient des	**écoles**	où les anciens enseignoient aux enfans ces
[3]	C002	, qui, au rapport de Suidas, tint une	**école**	de grammaire à Alexandrie, puis à Constantinople
[4]	C004	de différentes manieres dans les diverses	**Écoles**	de Peinture, sans cesser d'être le même ; le fond
[5]	C004	, dont on s'est servi jusqu'à présent dans les	**Écoles**	, pour nommer les parties de la frase ? Non, on est
[6]	C004	singularite que j'ai, abandonné ces termes de l'	**École**	; mais uniquement, parce qu'ils m'ont paru ne pas
[7]	C055	Non, sans doute. Il n'était plus ce temps où des	**écoles**	publiques, ouvertes et entretenues à grands
[8]	C061	disparut l'esprit exclusivement patricien de l'	**école**	classique, que commencèrent à s'introduire
[9]	C061	), globe, mode, proche (prŏpius), rose,	**école**	, sole (sŏlea), viole, v.-fr. voche (vŏco
[10]	C061	de o devant m et n : pomme, don, raison, bon,	**école**	. - 2) De u bref ou de y : trop (lat.moy. truppus

Abb. 11: Konkordanzen zur Abfrage: Lemma *école* (Lemma *Schule*) (Frantext)

Frantext ist lemmatisiert und morphosyntaktisch annotiert.

Andere Korpora mit Französisch

Ein großes Korpus mit französischen Internettexten ist **frWaC** mit 1,6 Mrd. Positionen. Dieses Korpus ist ebenfalls lemmatisiert und morphosyntaktisch annotiert. Es ist zugänglich und recherchierbar auch über KontText auf der Homepage des ČNK (2014).
Es gibt auch ein Korpus des gesprochenen Französischen (etwa 400.000 Wörter) aus unterschiedlichen frankofonen Regionen - **CRFP**.

4.1.3 Griechisch

Nationales Korpus Griechischer Sprache (EThEG)

Εθνικός Θησαυρός Ελληνικής Γλώσσας / Helenic National Corpus

http://hnc.ilsp.gr/en/
Es ist ein repräsentatives Korpus des gegenwärtigen Griechisch: 47 Millionen laufende Wörter, ausschließlich modernes (nach 1990) geschriebenes Griechisch, stilistisch ausgewogen. Das Korpus wird laufend erweitert.
Interface und Manual sind auf Griechisch und Englisch.

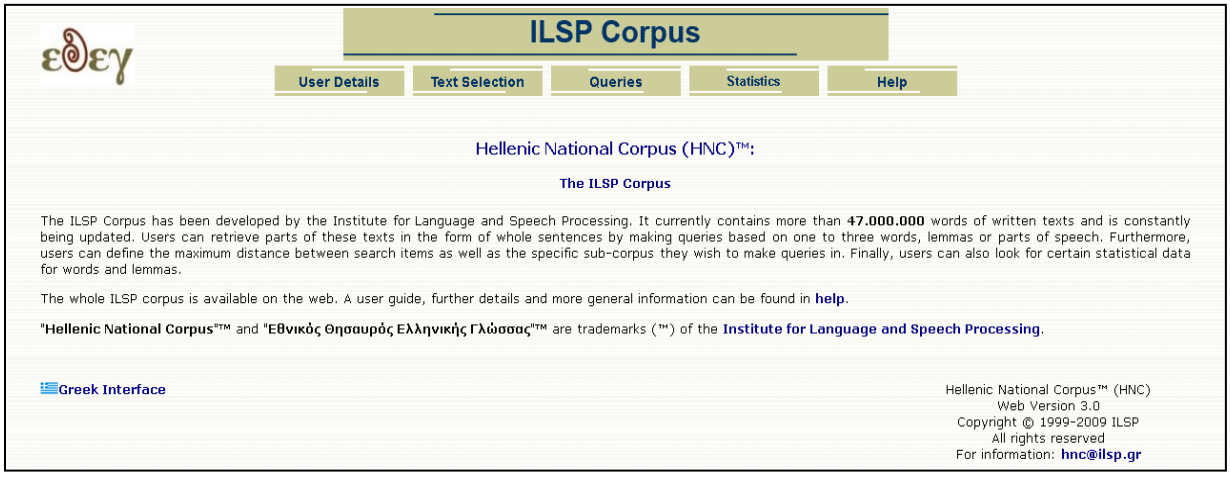

Abb. 12: EThEG/ HNC-Startseite (2014)

Abfrage
Queries → Lemma: **σχολείο**

1	ατος του καθηγητή φυσικής αγωγής σε όλα τα σχολεία.
2	επιστημονικές βιβλιοθήκες και σταδιακά τα σχολεία.
3	έχει προβλεφθεί η δυνατότητα καθιέρωσης σε σχολεία ανώτερης
4	αλογία 5 έως 15 χρηστών ανά υπολογιστή στα σχολεία έως το 2004, πράγμα
5	σης για απομακρυσμένα δημόσια σημεία, όπως σχολεία, κέντρα υγείας,

Abb. 13: Konkordanzen zur Abfrage: Lemma *σχολείο* (Lemma *Schule*) (EThEG/ HNC)

EThEG ist lemmatisiert und morphosyntaktisch annotiert.

4.1.4 Italienisch

Korpus des geschriebenen Italienisch(CORIS)

Corpus di italiano scritto / Corpus of written Italian

http://corpora.dslo.unibo.it
Das Korpus hat über 130 Mio. Wörter der geschriebenen Sprache, ist repräsentativ für das moderne Italienisch und stilistisch ausgewogen. Es wird sukzessive erweitert.
Die Homepage ist auf Italienisch und Englisch, das Abfrageformular auf Englisch.

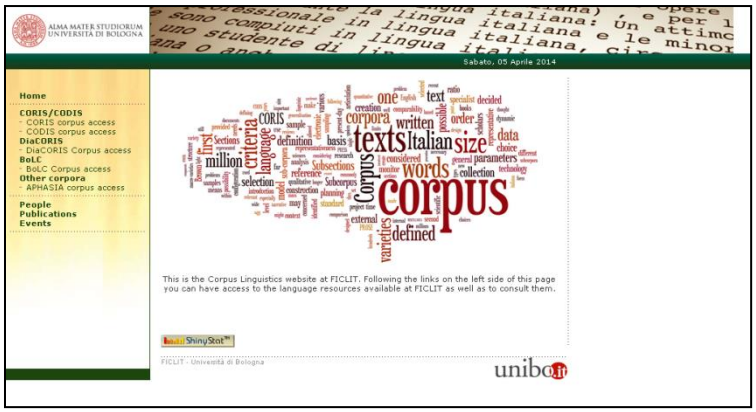

Abb. 14: CORIS-Startseite (2014)

Abfrage
Query: [lemma="scuola"]

Corpus query result		
Query: [lemma="scuola"]		
Sorted: 0		
Number of concordances: 30/30801		
, qualcuno all ' interno della	**scuola**	avesse saputo o intuito quant
ure indicate . Iscrizione alla	**scuola**	materna Per_quanto riguarda l
re_a innumerevoli parrocchie ,	**scuole**	, gruppi locali . Nata nel di
t La gastronomia italiana , la	**scuola**	di cucina , un elenco di rist
Papini con il suo Chiudiamo le	**scuole**	. Lo scrittore denunciava « l
alla fine del primo giorno di	**scuola**	, è parso raggiante . Ha subi
muro del convento adiacente la	**scuola**	e si mosse per prenderla . Po
né è mai stato il ruolo delle	**scuole**	di retorica nei secoli passat
non aveva mai avuto problemi a	**scuola**	; invece pensò come poteva fa
a ? Anche voi uomini avete una	**scuola**	? - soffia il Piccolo Drago s

Abb. 15: Konkordanzen zur Abfrage: Lemma *scuola* (Lemma *Schule*) (CORIS)

CORIS ist lemmatisiert und morphosyntaktisch annotiert. Die Abfragesyntax entspricht derjenigen, die im CQL-Mode im InterCorp verwendet wird. (Es gilt natürlich nicht für die morphosyntaktischen Zeichen (Tags).)

Andere Korpora mit Italienisch

Ein großes Korpus für Italienisch ist auch das "LA REPUBBLICA" CORPUS mit 380 Mio. Wörtern aus der gleichnamigen Tageszeitung oder das Korpus aus italienischen Internettexten (itWaC). Das Korpus **itWaC** ist gemeinsam mit dem Web-Korpus des Deutschen (deWaC), Französischen (frWaC) und des britischen Englisch (ukWaC) auch über die Homepage des ČNK (2014) zugänglich und recherchierbar.

4.1.5 Kroatisch

Kroatisches Nationalkorpus (HNK)

Hrvatski nacionalni korpus / Croatian National Corpus

http://www.hnk.ffzg.hr/
Direktlink zu NoSketch: http://filip.ffzg.hr/bonito2/run.cgi/first_form
Das Korpus hat d.Z. rund 115 Mio. laufende Wörter, wird regelmäßig ergänzt und ist stilistisch ausgewogen.
Zugang über Bonito oder NoSketch Engine. Das Interface und Manual sind auf Kroatisch und Englisch.

Abb. 16: HNK-Startseite (2014)

Abfrage
Query Type: Lemma → Lemma: **škola**

doc#11	s radovima djece iz osnovnih	**škola**	, uz tvrdnju da će ' svima onima koji su
doc#26	više od 700 svjedodžbi za srednje	**škole**	i fakultete ,
doc#30	u crkvi i u Hrvatskoj osnovnoj	**školi**	i gimnaziji u Budimpešti .
doc#67	na stranim sveučilištima ili srednjim	**školama**	, a određenom broju smo
doc#68	planova je i osnivanje posebne	**škole**	u Opatiji za obrazovanje managemanta .

Abb. 17: Konkordanzen zur Abfrage: Lemma *škola* (Lemma *Schule*) (HNK)

Das **HNK** ist lemmatisiert und morphosyntaktisch annotiert.

4.1.6 Polnisch

Nationalkorpus der Polnischen Sprache (NKJP)

Narodowy Korpus Języka Polskiego / National Corpus of Polish

http://nkjp.pl

Es ist ein stilistisch ausgewogenes Korpus mit 1.800 Millionen laufenden Wörtern der polnischen, ausschließlich geschriebenen Sprache (aufgeteilt in Presse-Texte und Bücher). Das Interface und eine Kurzbeschreibung sind auf Polnisch und Englisch.

Abb. 18: NPKJP – Startseite (2014)

Abfrage

Query/Zapytanie: **[base="szkoła"]**

- W dzień jest w	**szkole [szkoła:subst:sg:loc:f]**	, ale nie mogę jej
- Gdzie ona chodzi do	**szkoły [szkoła:subst:sg:gen:f]**	? - Tutaj, niedaleko
- A ja zaraz po	**szkole [szkoła:subst:sg:loc:f]**	. Uśmiechał się kiedy ruszyli
budynku z cyklu "1000	**szkół [szkoła:subst:pl:gen:f]**	na 1000-lecie". Ze
było go więcej w wiejskich	**szkołach [szkoła:subst:pl:loc:f]**	. W województwie łódzkim skrzyżowano
słyszała już takie nazwisko w	**szkole [szkoła:subst:sg:loc:f]**	, przyznała tę kwaterę właśnie
bitew podręczniki historii obowiązujące w	**szkołach [szkoła:subst:pl:loc:f]**	Obozu Pokoju przyznawały na skąpo
dzieci przepytywanych na lekcjach w	**szkole [szkoła:subst:sg:loc:f]**	, co też w domach
zebrał informacje o nastrojach w	**szkole [szkoła:subst:sg:loc:f]**	podstawowej im. "Walczącej
, ulic, budynków,	**szkół [szkoła:subst:pl:gen:f]**	, mostów, placów,

Abb. 19: Konkordanzen zur Abfrage: Lemma *szkoła* (Lemma *Schule*) (NPKJP)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert. Die Abfrage nach einem Grundwort (Lemma) liefert automatisch bei jedem KWIC auch morphologische Angaben (hier Wortart, Numerus, Kasus, Genus).

4.1.7 Russisch

Nationalkorpus der russischen Sprache (NKRJa)

Национальный корпус русского языка / Russian National Corpus

http://www.ruscorpora.ru/
Das Russische Nationalkorpus hat über 149 Mio. Tokens (entspricht etwa 100 Mio. Wörtern). Mehr als die Hälfte bilden Sachtexte, etwa 40% Belletristik und weniger als 4% gesprochene Sprache.
Die geschriebenen Texte stammen aus der Mitte des 18. Jahrhunderts bis heute, gesprochene Texte wurden zwischen 1950 und heute aufgenommen und für das Korpus aufbereitet.
Die Interface sowie eine gute Beschreibung des Korpus inkl. einiger statistischer Angaben zu den Texten und Tokens gibt es sowohl auf Russisch als auch auf Englisch.

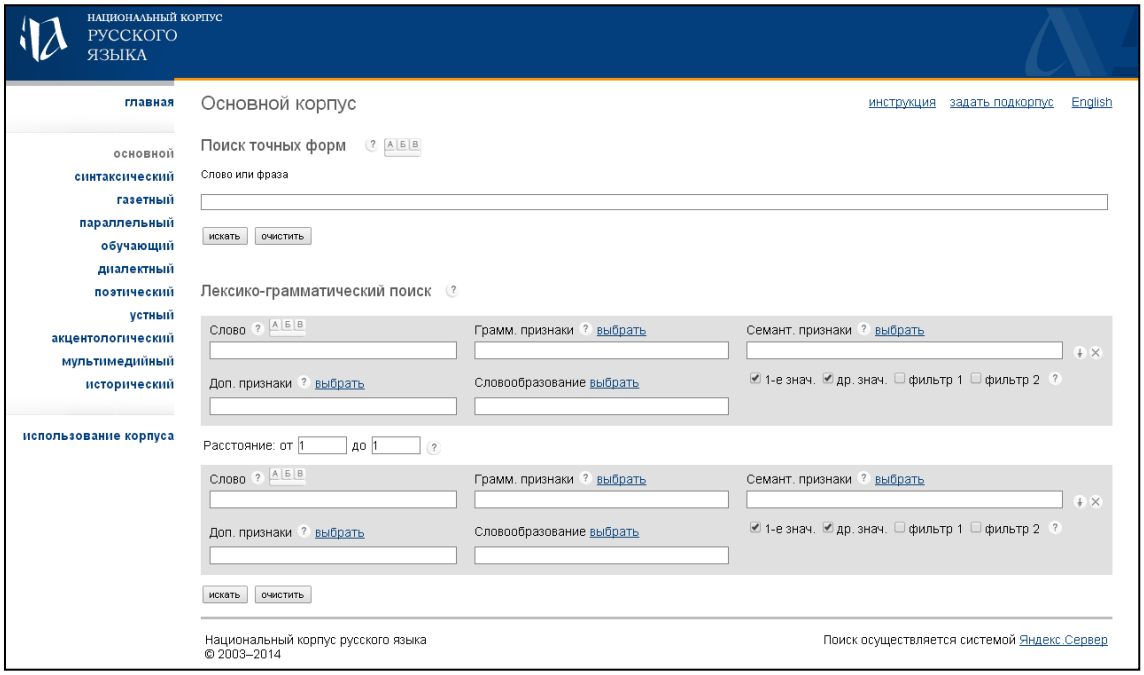

Abb. 20: NKRJa/ RNC – Abfrageseite (2014)

Abfrage
Слово/ Word[13]: **школа**

тебе понятие о той великой	школе	жизни, которая образует истинных человеков ←...→
03-06-00139а, и Програ́ммой подде́ржки нау́чных	школ	, грант N 005-97893. ←...→
Я же по́мню, как в	шко́ле	уви́жу таку́ю схе́му в кабине́те ←...→
вре́мя Ершо́в, впосле́дствии возгла́вивший программистскую	шко́лу	в Новосибирске, ещё рабо́тал в ←...→
в ча́стных ли́бо в воскре́сных	шко́лах	, "общеобразова́тельная же шко́ла не должна́ ←...→
при́нципами; напро́тив, оно́ изоби́лует разли́чными	шко́лами	, мно́гие из кото́рых тракту́ют одни́ ←...→
нау́к, профе́ссора Госуда́рственного университе́та - Вы́сшей	шко́лы	эконо́мики
пройду́т торже́ственные собра́ния, в подше́фных	шко́лах	— встре́чи с молодёжью, а в ←...→
Говори́шь, в	шко́ле	опя́ть ничего́ не за́дали? ←...→
день, когда́ Па́влик верну́лся из	шко́лы	, на́спех поел, кое-ка́к что́-то накалякал ←...→

Abb. 21: Konkordanzen zur Abfrage Lemma *школа* (Lemma *Schule*) (NKRJa/ RNC)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

[13] Word = Lemma bei Eingabe der Grundform

4.1.8 Slowenisch

Korpus der slowenischen Sprache (FIDAPLUS)

FIDAPLUS Korpus slovenskega jezika

http://www.fidaplus.net/
Korpus slovenskega jezika FIDAPLUS ist ein stilistisch ausgewogenes Korpus mit über 600 Millionen Textwörtern. Interface und Manual sind auf Slowenisch, Grundinformationen auch auf Englisch.

Abb. 22: FIDAPLUS-Startseite (2014)

Abfrage
Osnovno iskanje: **#1šola**

DELO..... 0000007	Škoda v novi gostinski	šoli	- Povodenj je v celjski občini hudo prizadejala vrsto šolskih
DELO..... 0000007	škode. Precej je poškodovana tudi nova stavba srednje gostinske	šole	in ob njej nova telovadnica. Na obeh objektih je
DELO..... 0000007	bo potrebno v celoti zamenjati, v kletni etaži nove	šole	pa so uničeni tlaki in stene. Na srečo ne
DELO..... 0000013	otrok v oddaljene kraje zaradi prostorske stiske v obeh črnomaljskih	šolah	in na pritiske občine in najemnika kotlarne na občane naselja
NOVI.TEDNIK. 0000005	mestni občini so se v projekt vključili dijaki Srednje ekonomske	šole	ter poslovna enota Telekoma Celje, katere ekipa je poskrbela
MENS.HEALT.. 0001116	izkaznico, prosim? Kje si ti, v srednji	šoli	? Leto dni je že od tega, sinko moj
MENS.HEALT.. 0001118	Kdo bo pospravil kopalnico, kdo bo peljal otroke v	šolo	, kdo se v družbi vede bolj prostaško, kdo
DNEVNIK..... 0000208	. Nagrade je prejelo tudi 15 dijakov z dolenjskih srednjih	šol	. Med nagrajenci so štirje mladi raziskovalci iz Ruske federacije
DNEVNIK..... 0000232	Učenci ribenske osnovne	šole	čakajo na šolski avtobus še vedno kar na cesti
DNEVNIK..... 0000234	Zaradi predvidene popotresne obnove	šole	učenci obiskujejo pouk na Bledu

Abb. 23: Konkordanzen zur Abfrage: Lemma *šola* (Lemma *Schule*) (FIDAPLUS)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

4.1.9 Slowakisch

Slowakisches Nationalkorpus (SNK)

Slovenský národný korpus / Slovak National Corpus

Die Version 6.1 public (aus Herbst 2013) beinhaltet geschriebene slowakische Texte aus den letzten Jahren. Deutlich überwiegen publizistische Texte (fast 70%), belletristische betragen etwa 14 %, Fachtexte 15%. Das Korpus 6.1 hat 655 Mio. Wörter.
Für registrierte Nutzer sind auch andere Korpora recherchierbar: Korpus der gesprochenen slowakischen Sprache, einige Web-Korpora, sowie (noch relativ kleine) Parallelkorpora Slowakisch – Bulgarisch, Englisch, Latein, Russisch, Tschechisch, Ungarisch.

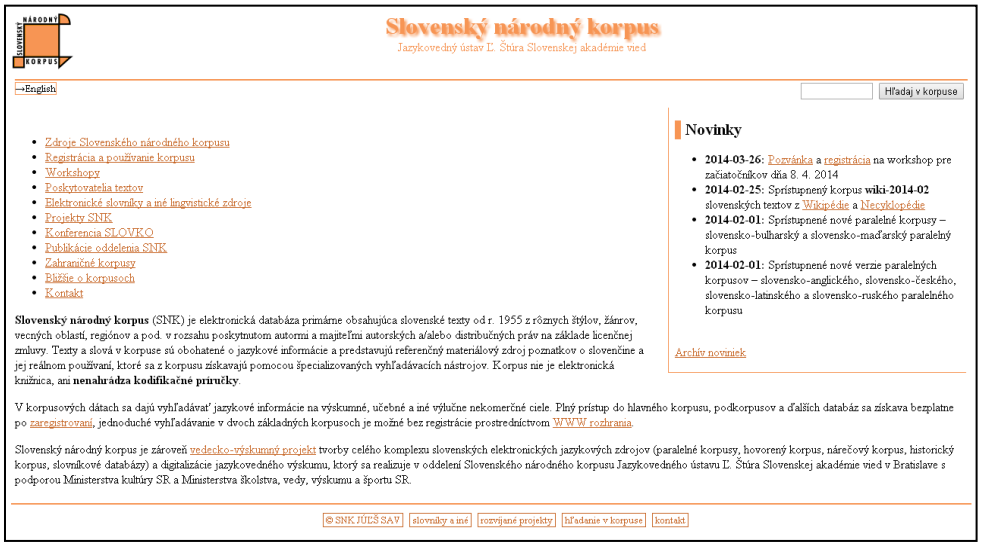

Abb. 24: SNK-Starseite (2014)

Abfrage

Query Type: Lemma → Lemma: **škola**

? Kňazi potrebujú poznať umenie . Po skončení	**školy**	prichádzajú do kostolov , o ktorých vedia len
Tvorivá dielňa Krása z hliny bude v základnej	**škole**	určená pre všetkých , ktorí si budú chcieť
potrebám chlapcov . Na systéme vyučovania na základných	**školách**	sa podpísala najmä ženská ruka , a preto
iniciátor projektu . Podľa neho sa deti v	**škole**	učia rozličné veci , ale mali by sa
ceslovenskej výtvarnej súťaži , ale výtvarná výchova v	**škole**	vraj nepatrí k jeho najobľúbenejším predmetom . „
19 . júla v areáli amfiteátra a základnej	**školy**	v Malatinej na Orave . Úvod sobotného programu
, a tak poskytli argumenty pre vznik ďalších	**škôl**	. Kvantitatívny rozvoj univerzít má však aj iný
sa narodil v Starom Smokovci . Do ľudovej	**školy**	chodil v Poprade , do gymnázia v Kežmarku
ôsmom , ani M . Horkheimer a frankfurtská	**škola**	. V Nemecku siaha tradícia myslenia o politike
že nedokáže pokračovať v štúdiu ani na inej	**škole**	. Určitá časť šikanovaných vidí v samovražde jediné

Abb. 25 Konkordanzen zur Abfrage: Lemma *škola* (Lemma *Schule*) (SNK)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

4.1.10 Spanisch:

Korpus des Spanischen/ CORPUS DEL ESPAÑOL

http://www.corpusdelespanol.org/
Das Korpus beinhaltet Texte ab dem Jahr 1200 bis heute, inkludiert auch Texte gesprochener Sprache aus dem 20. und 21. Jahrhundert.
Manual und Interface sind auf Englisch.

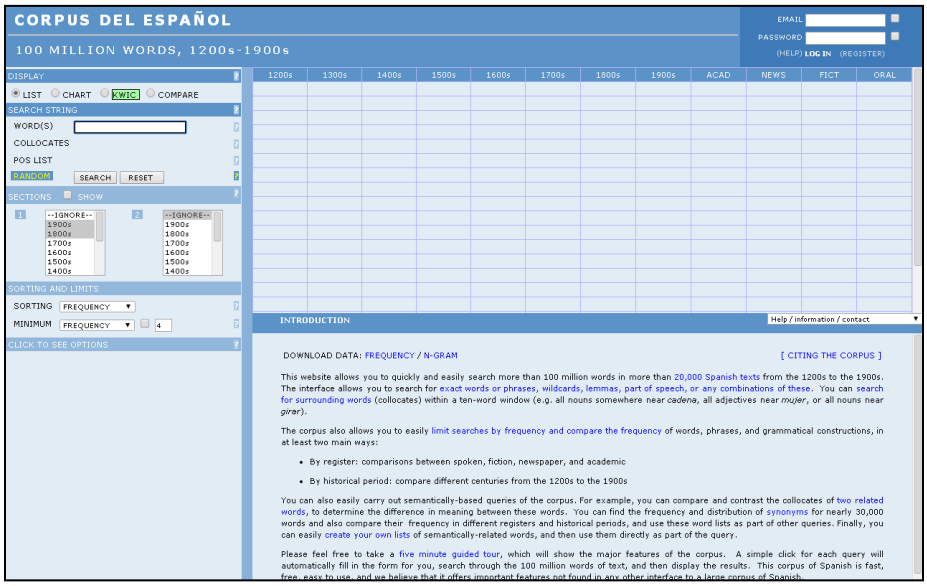

Abb. 26: Corpus del Español - Start- und Abfrageseite (2014)

Abfrage
WORD(S): **[escuela]**

por la que agora tengo . Juan . - ¿ Tienen	escuelas	**allá** ? Pedro . - Infinitas . Los señores , y primera
" . Se considera que la etapa culminante de la "	escuela	**bolera** " discurre entre 1835 y 1880 , en coincide
Massachusetts (MIT)- - o del norte de California donde hay	escuelas	**como** Stanford , que producen una corriente con
en lo que tienen de fundamental y común a todas las	escuelas	**conservadoras** ? ¿ No es esto lo que he hecho a
conocen á Cristo ; los que nunca han entrado en la	escuela	**de** Cristo no se puede decir que niegan á Cristo .
en primaria . Pero después teníamos en nuestra universidad , la	escuela	**de** educación , que preparaba a los pedagogos ,
en esta parte más novedad que la de franquear también estas	escuelas	**de** toda pensión o retribución particular . Cabalm
frecuentes en los porteros de las mismas casas . Si la	escuela	**debe** ser para el niño un lugar sano y fortificante
gran neoconfuciano Zhu XI , que desarrolló las doctrinas de la	escuela	**del** principio . En el siglo XIV estas doctrinas fuer
haréis más llevadera , y convertiréis vuestra desgracia en una	escuela	**donde** habréis aprendido que el hombre para se
la subvención anual (grant) votada en favor de las	escuelas	**elementales** , el mismo Departamento creó cier
a escuelas prácticas para agricultura , pero verdaderamente	escuelas	**eminentemente** prácticas , no impartir la enseñ
y seudo - filosófico , comunes a las diferentes sectas y	escuelas	**en** que se divide y s

Abb. 27: Konkordanzen zur Abfrage *escuela* (Lemma *Schule*) (Corpus del Español)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

Von demselben Autor (Mark Davies), mit der gleichen Schnittstelle und mit gleichen Recherchemöglichkeiten gibt es auf der Homepage http://corpus.byu.edu/corpora.asp Links zu anderen hilfreichen Korpora: die meisten sind zu (amerikanischem, kanadischem und britischem) Englisch, aber auch ein relativ großes Korpus (45 Mio. Wörter) zum Portugiesischen.

4.1.11 Tschechisch

Tschechisches Nationalkorpus (ČNK)

Český národní korpus/ Czech National Corpus

https://www.korpus.cz/
Das Instrument besteht aus mehreren Korpora. Im synchronen Teil der geschriebenen Sprache befinden sich mehr als 2 Mrd. Wörter. Die gesprochene tschechische Sprache wird mit Transkripten von fast 3 Mio. Wörtern repräsentiert, zu ihnen ist auch immer die Tonspur abrufbar.
Historische tschechische Texte sind im Diakorp zu finden (fast 5 Mio. Wörter).
Über die Schnittstelle KonText des Tschechischen Nationalkorpus gelangen registrierte Nutzer auch zu großen Korpora anderer Sprachen und auch zum InterCorp.
Das Interface und die Beschreibungen sind auf Tschechisch und Englisch.

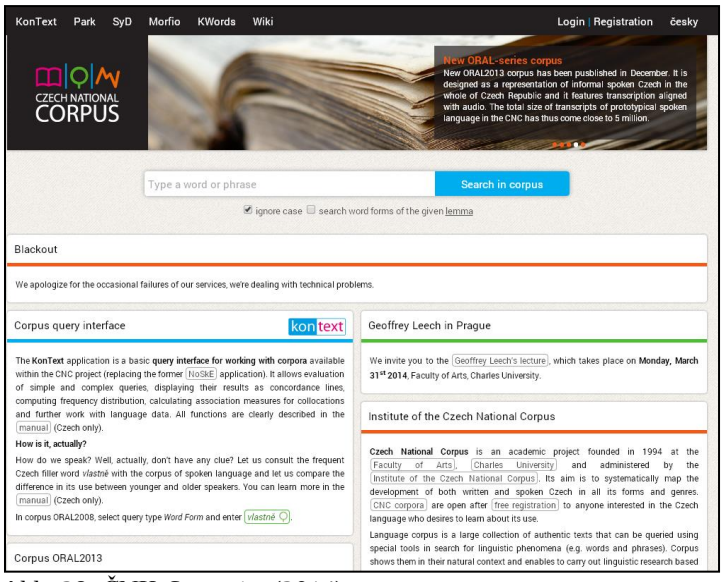

Abb. 28: ČNK-Startseite (2014)

Abfrage
Query Type: Lemma → Lemma: **škola**

# Corpus: syn			
# Hits: 2071691			
# Relative frequency: 771.54 ((related to the whole syn))			
# ARF: 663956.25			
# Query: word,[lemma="škola"] 2071691			
# Shuffle: enabled			
# GDEX: 40			
1	ıe . A pak mi řekl : Já vím , že jste to ve	**škole**	ještě neměli . Tomu se budete učit , až trochu povyrosteš . Ale můžeš panu
2	a většina z nich studovala na vysokých	**školách**	v USA . Tony se mi moc snažil vysvětlit ten jejich svět . Nakonec jsem
3	vidět malá děvčátka , když jdou ráno do	**školy**	. Každé je vyparáděno a mašle ve vlasech jim nikdy nechybí . Není tu vlastně
4	klá ze svého povolání učitelky v hudební	**škole**	. A neříká se , že každá učitelka je tak trochu praštěná ? " Kime
5	vě se tento den dokonce zavírají střední	**školy**	, protože by stejně skoro nikdo nepřišel . Počítá se s tím , že první
6	Pokrok asi takový , jako když žák lidové	**školy**	umění při hře na housle nepřetrhne strunu . Fořt Bořivoj , když se tu novinu
7	stli jste , zmužněli . Když jste byli ještě	**školou**	povinní , vypadali jste jinak . " Bojovníci za přírodu stáli přede dveřmi ředitelova domku
8	n , skoro šedesátník , představitel staré	**školy**	, si ještě pamatoval přísloví " dlouhé vlasy , krátký rozum " . Snad právě
10	vitý klacek , co ani nedochodil základní	**školu**	, plivl ze vzteku do talíře ? " Ingmar odstrčil snídani a přivolal servírku .
20	ryb (anglicky se říká school of fish - čili	**škola**	ryb) , což není docela lehké , je - li vítr . Ale tenkrát

Abb. 29: Konkordanzen zur Abfrage: Lemma *škola* (Lemma *Schule*) (ČNK)

Das Korpus ist lemmatisiert und morphosyntaktisch annotiert.

4.1.12 Türkisch

Türkisches Nationalkorpus (TUD)

Türkçe Ulusal Derlemi / Turkish National Corpus

http://www.tnc.org.tr/index.php/en/

Das Türkische Nationalkorpus beinhaltet etwa 50 Millionen Wörter in Texten aus den Jahren 1990 – 2009, etwa 2% bilden Transkripte der gesprochenen Sprache. Es ist stilistisch ausgewogen. Das Interface und die Beschreibung sind auf Türkisch und Englisch.

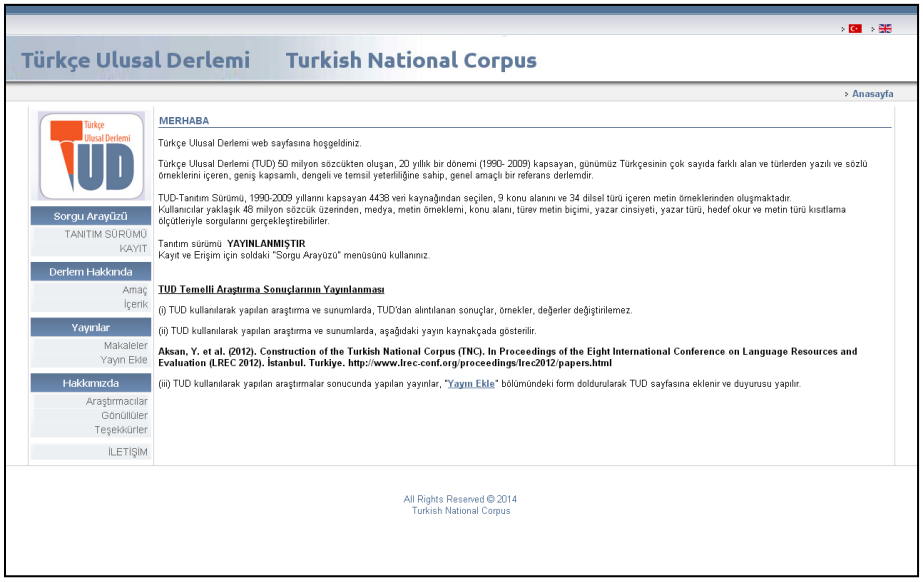

Abb. 30: TNC-Startseite (2014)

Abfrage
Query Term: **okul***

1	konuştu; - "Hocam, biz hedefteyiz.	**Okuldan**	ayrılmak istiyoruz. Biliyorsunuz dün
2	göreceğimiz çok açıktır. Bunun için	**okullarda,**	Prof. Kalter'in dediği gibi mesela
3	rahibeler yılar. Romana konu olan	**okul**	150 yıl önce açılan Özel
4	içindi senin özbenliğinin. Bak baba,	**okula**	iki yıl erken başlatılmış her
5	Devlet Üretme Çiftlikleri, tarım meslek	**okulları,**	tohum ve arazi ıslah istasyonları
6	muhit, gittiği yazlık yer, okuduğu	**okul,**	babasının işi, kıyafetlerinin markası, taktığı
7	zekâlı bedevi, kuş beyniyle bizim	**okula**	nasıl birincilikle girdi, anlamıyorum. Oğlu
8	Enstitüsü mezunu, daha sonra o	**okulun**	adı Öğretmen Okulu olarak değiştirilmiş,
9	tılarımızın ortasında Ali'yle gelmiştik okula?	**Okulun**	o büyük konferans salonunda kura
10	miyorlardı. İşleriyle öylesine doluydular ki!	**Okulda**	yoruldukları yetmiyor gibi çalışmalarını, e

Abb. 31: Konkordanzen zur Abfrage: Wortform *okul-* (Wortform *Schule* mit offenem Ende) (TUD)

Das TUD ist (noch) nicht morphosyntaktisch annotiert und lemmatisiert.

4.1.13 Ungarisch

Das ungarische Nationalkorpus (MNSz)

Magyar Nemzeti Szövegtár / Hungarian National Corpus

http://corpus.nytud.hu/mnsz/

Das Ungarische Nationalkorpus umfasst fast 188 Mio. laufender Wörter in Texten aus den letzten Jahrzehnten. Fast die Hälfte der Texte stammt aus Presse und Publizistik. Neben der üblichen Recherche (nach unterschiedlichen Attributen) ist auch die Eingrenzung der Region möglich, dadurch können Varianten des Ungarischen (aus Ungarn, der Slowakei, Transkarpatien, Transsylvanien und Vojvodina) verglichen werden.

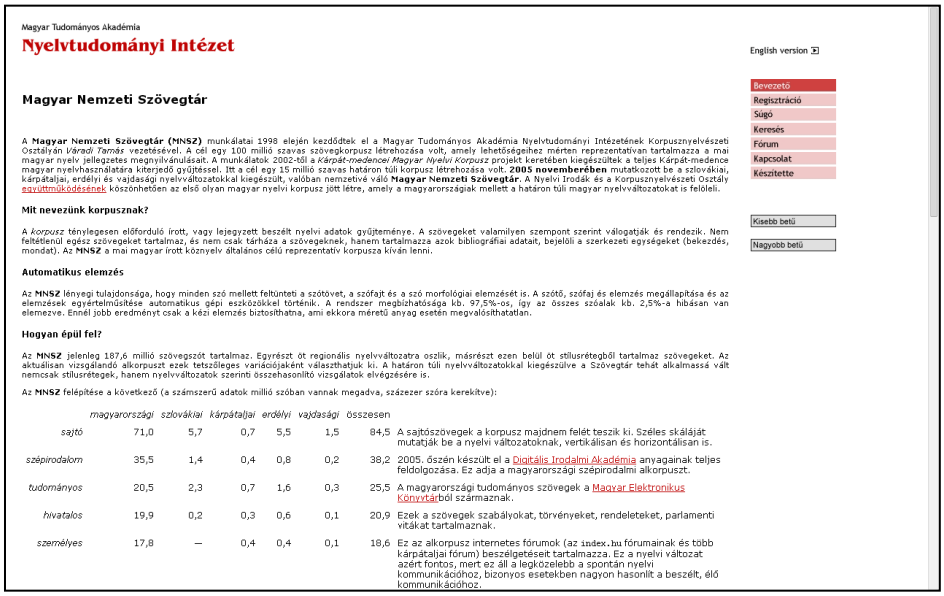

Abb. 32: MNSz-Startseite (2014)

Abfrage

Query → Stem: **iskola**

Region: Hungary, Slovakia, Subcarpathia, Transylvania, Vojvodina			
Subcorpus: media, literature, scientific, official, informal			
Query: [lemma = "(.*|)?iskola(|.*)?"] ;			
Size of corpus queried: 187644886 words			
Number of matches: 64094 341,57 / million words			
1.	Az erõs akaratú Márton István	**iskolája** *N.PSe3.NOM*	1804-től, a református konvent
2.	: 1. óvodákra 2. általános	**iskolákra** *N.PL.SUB*	3. gimnáziumokra, szakközépiskolákra,
4.	a cél, hogy az	**iskoláknak** *N.PL.DAT*	elvegyék a kedvét a mérlegképes
5.	, hogy az egyház és	**iskolája** *N.PSe3.NOM*	felfelé ívelõ pályára kerüljön.
6.	hogy a hibát nemcsak az	**iskolában** *N.INE*	kellene keresnie? Azt állítja
7.	betakarítási munkák, egy hét	**iskola** *N.NOM*	közben / helyett; káposzta
14.	. Akik Zsonnát ismerték az	**iskolából** *N.ELA*	vagy Bísztõ Gyalla és Karmenják
15.	hogy világnézeti beállítottság nélkül nincs	**iskola** *N.NOM*	, ahogy nincs világnézet nélkül
16.	, a minõség iránt fogékony	**iskolákról** *N.PL.DEL*	, és a fasornak talán
18.	óvodások napi szállítása a központi	**iskolába** *N.ILL*	, mert annyira lecsökkentették a
20.	jövõje van egy ilyen kicsi	**iskolának** *N.DAT*	? Nem kerülnek -e hátrányba

Abb. 33: Konkordanzen zur Abfrage: Lemma *iskola* (Lemma *Schule*) (MNSz)

Das MNSz ist lemmatisiert und morphosyntaktisch annotiert.

Schlussbemerkung zu nationalen Korpora

Neben den vorhin erwähnten Korpora gibt es im Internet eine Menge computerlesbarer Textsammlungen, die auch (zu Recht) als Korpora bezeichnet und zur Verfügung gestellt werden. Für ihre Verwendung als vollwertige Sprachkorpora benötigt man jedoch einen Korpusmanager, was die Arbeit wegen der aufwendigen Aufbereitung erschwert und für viele potenzielle Nutzer uninteressant macht.

Es fehlen aber noch jegliche Korpora mit Sprachen vieler Minderheiten: Romani, Vietnamesisch, Kurdisch, um einige zu nennen. Für das Türkische gibt es zwar ein Nationalkorpus, ein deutsch-türkisches Parallelkorpus wird erst im Rahmen des Projektes InterCorp geplant.

4.2 KORPORA mit DEUTSCH

Die deutsche Sprache ist in verschiedenen Korpusinstrumenten gut erfasst. Einzelne Institute, die sich mit der Korpuserstellung befassen, finden sich im ganzen deutschsprachigen Raum. Da aber unter den Instituten kaum eine Zusammenarbeit besteht, sind die Schnittstellen, Korpustools, Eigenschaften und Abfragemodi bei jedem Korpus anders, was für ihre Nutzer gewisse Schwierigkeiten mit sich bringt. Bei der Arbeit mit Korpora aus verschiedenen Institutionen ist also eine große Flexibilität gefragt.

Der Umgang mit jedem Korpusmanager braucht Routine. Viele Eigenschaften bzw. Abfragemöglichkeiten sind oft in einer eigenen Korpussprache verschlüsselt, Beschreibungen irgendwo auf der Homepage versteckt oder gar nicht zu finden. Die folgende Übersicht soll die Arbeit mit den Korpora erleichtern, die Eigenschaften kurz zusammenfassen und die wichtigsten Abfragen (inkl. einiger Rechercheergebnisse) darstellen.

Die Reihenfolge, in der die einzelnen Korpora vorgestellt werden, wurde aufgrund der Komplexität der Instrumente vorgenommen (dies ist natürlich auch subjektiv). Erfahrungsgemäß ist es ratsam, mit einem einfachen Instrument zu beginnen, deswegen wird mit dem **Wortschatz-Portal** der Universität Leipzig angefangen, danach folgt das **DWDS** (Berlin-Brandenburgische Akademie der Wissenschaften) und das im Aufbau befindliche plurizentrische Projekt **Korpus-C4**.

Die größten und wohl auch komplexesten Korpora für Deutsch sind in Mannheim zu Hause: **Das Deutsche Referenz Korpus** für die geschriebene Sprache und die **Datenbank für gesprochenes Deutsch**. Die Übersicht wird um Korpora mit historischen Texten ergänzt.
Es gibt natürlich auch noch andere hilfreiche Korpora für Deutsch, die hier aber nicht näher erwähnt werden: Die meisten von ihnen (**deWaC**, **AAC**, **Korpus Südtirol**, **Schweizer Textkorpus**) sind in andere Projekte oder Großkorpora integriert, weitere sind erschwert zugänglich (TenTen corpora[14]), bzw. für DaF/DaZ nur wenig relevant.

[14] Webkorpora mit mindestens einer Billion (10^{10}) Wörtern. Ein großer Vorteil dieses Korpus ist das Tagging mit RFTAgger. Dieser ermöglicht auch das Suchen nach Kasus, Genus und anderen morphosyntaktischen Kategorien, die über andere Tagger nicht abrufbar sind (vgl. Schmid/ Laws 2008).

4.2.1 Korpora mit (mehrheitlich) geschriebenem Deutsch

4.2.1.1 Wortschatz Leipzig

http://wortschatz.uni-leipzig.de/
Dieses Instrument verbindet ein elektronisches Wörterbuch mit einem großen Textkorpus. Zu den abgefragten Elementen werden morphologische, syntaktische und lexikalische Angaben angezeigt.
Das Korpus ist lemmatisiert – den einzelnen Wortformen werden Grundformen zugeordnet, im Ergebnis werden alle Formen aufgelistet. In den Konkordanzen erscheint jedoch als KWIC nur die abgefragte Form.
Es ist ein automatisch und „opportunistisch erstelltes Korpus“ (wie seine Autoren schreiben), daher soll man die Recherchergebnisse mit Vorsicht betrachten.

Abfragen:
Die Erstellung einer Abfrage ist sehr einfach. Ins Suchfeld schreibt man bloß das gewünschte Wort oder die gewünschte Verbindung. Andere Einstellungen sind nicht notwendig.
Da es sich primär um ein Instrument für die Erschließung des Wortschatzes handelt, sind die Abfragemöglichkeiten dementsprechend beschränkt. Grundsätzlich können nur Wortformen (mit der Möglichkeit Groß-/Kleinschreibung zu beachten) abgefragt werden. Wortkombinationen sind nur in der Abfolge abfragbar, in der sie ins Suchfeld eingegeben werden.

Abfragemöglichkeiten im Wortschatz – Leipzig
Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis (Beschreibung)
Arzt	*Arzt* (nur diese Wortform)
Arzt*	*Arzt*, *Arztabrechnung*, *Arztakten*, *Arztangebot* ... (automatische Ordnung nach Formen, alphabetisch absteigend)
***arzt**	*Notarzt*, *Chefarzt*, *Hausarzt*, *Zahnarzt* ... (automatische Ordnung nach Frequenz, absteigend)
böhmisch*·Dorf*	*böhmisches Dorf*, *böhmische Dörfer*

Tab. 2: Abfragen Wortschatz-Portal

Eingabe der Abfrage
Beispiel: Wortform *Hörer*

- **Einstellungen:** keine erforderlich
- **Abfrage** (Suchfeldeingabe): **Hörer**

Ergebnis: Angaben und Belege in Form einer Tabelle (Abb. 34) mit morphologischen, grammatischen und semantischen Angaben, sowie Belegen (Kontext von mindestens einem Satz) und Listen mit (automatisch errechneten) Kollokationspartnern.

Ergebnis der Abfrage *Hörer*:

Wort: Hörer
Anzahl: 3128
Häufigkeitsklasse: 12 (d.h. *der* ist ca. 2^12 mal häufiger als das gesuchte Wort)
Sachgebiet: Nachname
Allgemeines
Maschinen
Morphologie: hör|er
hör|er
Grammatikangaben: Wortart: Substantiv
Wortart: Eigenname
Geschlecht: männlich
Flexion: der Hörer, des Hörers, dem Hörer, den Hörer
die Hörer, der Hörer, den Hörern, die Hörer
Relationen zu anderen Wörtern:

- Synonyme: Kommilitone, Student, Studierende, Studiosus, Telefonhörer, Zuhörer
- ist Synonym von: Empfänger, Lernender, Student
- wird referenziert von: Schüler

Links zu anderen Wörtern:

- Grundform: Hörer
- Teilwort von: Hörer auflegen, den Hörer auflegen
- Form(en): Hörer, Hörern, Hörers
- Weibl. Form: Hörerin

Beispiel(e):
Mit diesem akustischen Wiedererkennungsmerkmal wissen die Hörerinnen und **Hörer** sofort, dass sie den richtigen Sender eingestellt haben. (Quelle: *www.radio.li, 2011-01-20*)
«Die Songs sind eine Herausforderung für die **Hörer**», sagt Dominik Deuber. (Quelle: *www.stadi-online.ch, 2011-01-25*)
weitere Beispiele
Signifikante Kookkurrenzen für Hörer:
Hörerinnen (7295.49), Radio (1655.77), Bremen (1008.11), Bremen Vier (822.64), etc.
Signifikante linke Nachbarn von Hörer:
und (1199.32), den (1149.59), die (689.22), zum (437.41), seine (412.38), etc.

Abb. 34: Abfrage *Hörer* (Wortschatz-Portal)

Die Graphische Darstellung kann sehr schnell feste Verbindungen aufdecken – diese sind bei anderen Korpusinstrumenten eher schwierig zu identifizieren. Die häufigsten Kollokatoren/ Nachbarn/ Kookkurrenten werden auch in Form einer Graphik („Spinnennetz“) angezeigt:

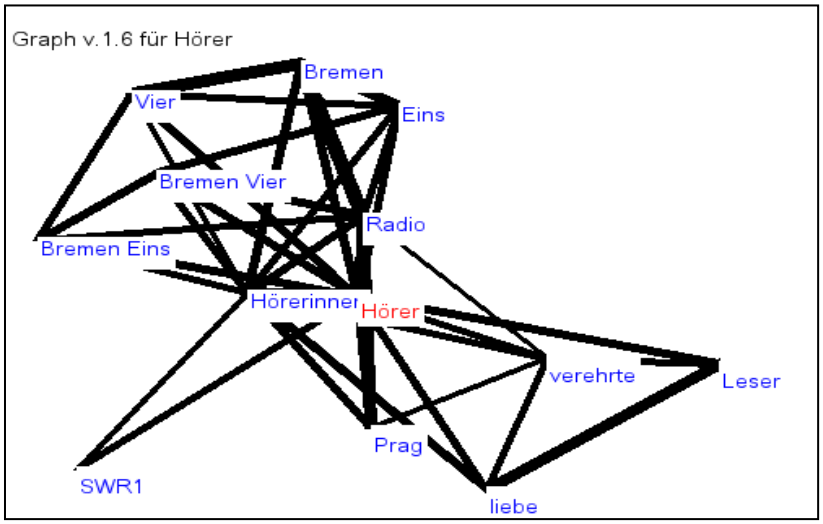

Abb. 35: Graphische Darstellung der Kollokatoren zur Abfrage *Hörer* (Wortschatz-Portal)

Aus der Graphik ist ersichtlich, dass das Wort *Hörer* selten ohne die movierte Form *Hörerin* vorkommt. Darüber hinaus sieht man deutlich, dass das Wort *Hörer* im heutigen Deutsch vorrangig in der Bedeutung *Radio-Zuhörer* verwendet wird und viel seltener als Synonym zu *Student* oder als *Kopfhörer (eines Telefons).*

Auf dem **Wortschatz-Portal** sind auch Wörterbücher von anderen 230 Sprachen. Alle wurden automatisch erstellt, das Korpus besteht aus Internetseiten der jeweiligen sprachlichen Domänen. Dieses Instrument eignet sich gut zur Erklärung einiger Basisfunktionen von Korpora: Darstellung der Belege, Errechnung von Frequenzen, Ermittlung der Kollokationspartner, Erschließung des Kontextes (Umfelds) u.a. Der besondere Wert des **Wortschatz-Portals** liegt in der übersichtlichen und graphisch gelungenen Präsentation der Ergebnisse.

4.2.1.2 Das Digitale Wörterbuch der deutschen Sprache (DWDS)

www.dwds.de
Das Das Digitale Wörterbuch der deutschen Sprache (DWDS) sind eigentlich mehrere (Sub)Korpora. Die Texte stammen aus Zeitungen, Belletristik und Gebrauchsliteratur. Das Gros bilden Texte aus dem 20. Jahrhundert, welche stilistisch ausgewogen sind.
Das DWDS erfordert eine Anmeldung, mit der man Zugang zum größten Teil der Texte erhält.

Die Eigenschaften sind unter Ressourcen (rechts oben) sehr gut beschrieben.
Ins Suchfeld (nach der ersten Abfrage erscheint es links oben) wird die Abfrage eingegeben (das gewünschte Wort/ Wortteil/ Verbindung etc.), rechts davon wählt man die Ansicht (Standard-, Wörterbuchansicht etc.).

In der Standard- und in der Wörterbuchansicht kann man auch die Aussprache (allerdings nur bei Abfrage einzelner Wörter) anhören. Die Tonspur für einzelne Wörter haben professionelle Sprecher/-innen besprochen. Es handelt sich also um kein Korpus der gesprochenen Sprache! Korpora mit Vertonung können beispielsweise als phonetische Hilfe im Unterricht eingesetzt werden. Da es sich um ein Korpus mit ausschließlich bundesdeutschen Texten handelt, befolgt die Aussprache auch die bundesdeutsche Aussprachennorm (siehe dazu auch Kap. 5).

Die Syntax der häufigsten Abfragen befindet sich unter dem Button Hilfe zur Suche (auf der Homepage oben rechts) und wird auch hier zusammengefasst und ergänzt:

Abfragemöglichkeiten im DWDS
Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis (Beschreibung)
Arzt	*Arzt*, *Arztes*, *Ärzte* … (alle flektierten Formen von *Arzt*)
@Arzt	*Arzt* (nur diese Wortform)
Arzt*	*Arzt*, *Arztbesuch*, *Arztberuf* … (mit *Arzt-* beginnende Wörter)
***arzt**	*Sportarzt*, *Hausarzt* … (auf *-arzt* endende Wörter)
"gut·Arzt"	*guter Arzt*, *bessere Arzt*, *bester Arzt*, *besten Arztes*, *gute Ärzte* … (beide Wörter lemmatisiert)

/arzt/	*Zahnarzt, Oberarzt, knarzt, verwarzt, verharzt* … (alle Formen mit der Buchstabenkombination *arzt*)
$p=V*·with·/arzt/	*geharzt, knarzt, knarzten, verharzt, vewarzt* (alle Verben mit der Buchstabenkombination *arzt*)
"das·gute·Beispiel"	*das gute Beispiel, das beste Beispiel, die besseren Beispiele* … (alle Formen aller Wörter)
"Kanzler·#1·Schröder"	***Kanzler Schröder***, ***Kanzler*** *Gerhard* ***Schröder*** … (*Kanzler* und *Schröder* im Abstand von höchstens einem Wort)
"Kanzler·#3·Schröder"	***Kanzler*** *aller Moleküle Gerhard* ***Schröder*** *will Innovationen fördern;* ***Kanzler*** *und begnadeten Wahlmatador* ***Schröder*** … (*Kanzler* und *Schröder* im Abstand von höchstens drei Wörtern)
Kanzler·||·Schröder	Alle Sätze, in denen entweder das Wort *Kanzler* oder *Schröder* oder beide Wörter vorkommen.
Kanzler·&&·!Schröder	Alle Sätze, in denen *Kanzler*, aber nicht *Schröder* vorkommen.
$p=NE·with·Herzog	*Roman Herzog*, *Peter Herzog* … (Eigennamen[15] *Herzog*)
"Einfluss·#2·$p=NE"	*Einfluß Smetanas*, *Einfluss in der EU* … (*Einfluss(ß)* im Abstand von max. 2 Positionen von einem Eigennamen[15])
$p=NN·with·*lein	*Häuflein, Kräutlein, Büchlein* … (Substantive[15], die auf -*lein* enden. Ausgeschlossen.)
$p=ADJA·with·*sam	*schweigsam, ehrsam, wundersam, achtsam* … (Adjektive[15]auf -*sam*)
$p=ADJA·with·mini*	*minimal*, *minimalistisch*, *miniaturhaft* … (Adjektive[15] auf *mini*-)
"gehen·#5·aus·with·$p=PTKVZ"	*Das Licht* ***ging aus*** *(…), Mir* ***ging*** *die Luft* ***aus***. *Wir* ***gehen*** *also nicht von der Sprechhandlung* ***aus*** *(…)* … (Verb[15] *ausgehen* mit abgetrenntem *aus*, maximaler Abstand 5 Positionen)
"sein·with·$p=VVFIN·#20·$p=VVPP·#0·@worden"	
	… im September 1980 ***waren*** *Teile des Streckennetzes* ***stillgelegt worden***; *das anzuerkennen* ***sei*** *auf Grund der längst überholten These von der Einheit eines Volkes unmöglich* ***gemacht worden***; *jedoch* ***waren*** *sie gleichzeitig darauf aufmerksam* ***gemacht worden*** (Verb *sein* als finites Verb[15], max. 20 Positionen davon Partizipium eines Vollverbs gefolgt von *worden*)
\,	, (Suche nach Komma. Satzzeichen wie *",.?!;* werden mit Backslash maskiert!)

Tab. 3: Abfragen DWDS

Weitere Abfragemöglichkeiten sind unter DWDS → Hilfe zur Suche zu finden.
Das Korpus ist mit dem Stuttgarter Tagset (STTS) morphologisch getagget und lemmatisiert. Dasselbe Tagset findet man in den meisten Korpora der deutschen Sprache (DeReKo, deutsche Parallele des InterCorp, deWaC).

[15] Andere abfragbare morphosyntaktische Kategorien entsprechen dem Stuttgarter Tagset (STTS). Sie sind im Kap. 8.2 angeführt.

Eingabe der Abfrage
Vor der Abfrage sind keine Einstellungen notwendig. Die Ansicht (Default: DWDS-Standardansicht) kann auch nach Abrufen der Ergebnisse geändert werden.

Beispiel: Wortform *Hit*

- **Einstellungen:** keine erforderlich
- **Abfrage** (Suchfeldeingabe): **Hit**

Ergebnis: Je nach Ansichteinstellung erscheinen Fenster mit verschiedenen Angaben.

DWDS-Wörterbuch:

Hit
mask., -s, -s
Herkunft Englisch
Erfolgsschlager, Spitzenschlager
eine Schallplatte mit den neuesten Hits kaufen
Dazu: Hitparade

Abb. 36: Fenster: DWDS-Wörterbuch

DWDS-Wortprofil 3.0

"Rank	Stammform	Wortart	logDice	Frequenz
"1	Medley	Substantiv	11	18"
"2	Coverversioner	Substantiv	11	17"
"3	Oldies	Substantiv	11	44"
"4	Song	Substantiv	11	39"
"5	wie Lady	PP	11	6"
"6	Evergreens	Substantiv	10	18"
"7	Flops	Substantiv	10	25"
"8	für Kids	PP	10	50"
"9	Lied	Substantiv	10	33"
"10	Tantiemen	Substantiv	9	6"
"11	landete	Verb	9	237"

Abb. 37: Fenster: DWDS-Wortprofil – Präpositionalgruppe

DWDS-Kernkorpus 20

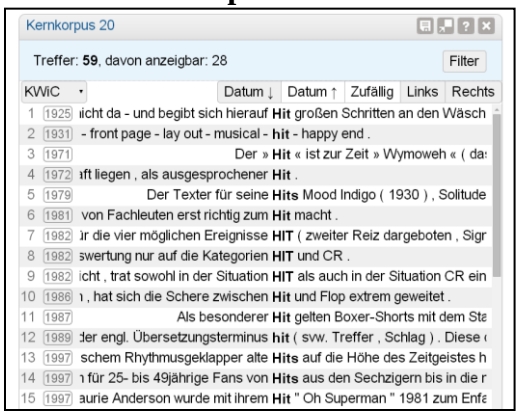

Abb. 38: Fenster: DWDS-Konkordanzen

Wortverlauf (Basis DWDS-Kernkorpus)

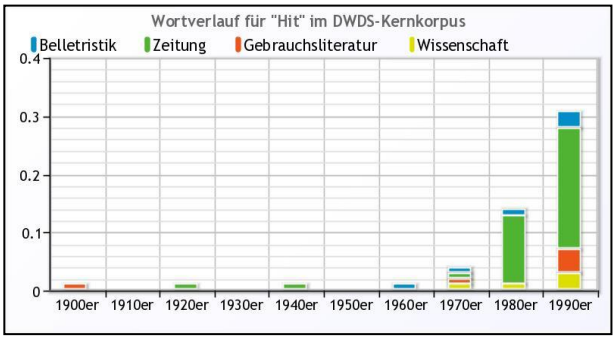

Abb. 39: Fenster: DWDS-Kernkorpus - Wortverlauf

In weiteren Fenstern der Standardansicht gibt es Angaben zu **Etymologie**, **Open Thesaurus**, **Konkordanzen** aus einzelnen Korpora (Zeit-Korpus, Kern-Korpus etc.) und das aus den Konkordanzen errechnete **Wortprofil** mit Angaben, die das Wort – je nach seiner Wortart - in seinen typischen Funktionen und Positionen im Satz zeigen: Überblick (Kookkurrenzen), Attribute zum Wort (falls vorhanden), Präpositionalgruppen, in denen das Wort vorkommt, Subjekte und Objekte, die das Wort (falls ein Verb) regiert, und andere mehr. Im Prinzip sind diese Informationen ident mit sog. Word Sketches – automatisch berechneten Eigenschaften der Wörter (siehe Sketch Engine 2014 und Kilgariff et al. 2014).

Es ist auch ratsam, andere Ansichten zu aktivieren, um sich ein Bild zu verschaffen, welche Informationen der Korpusmanager errechnen kann.

Das **DWDS** ist ein relativ einfaches Korpusinstrument mit umfangreichen Möglichkeiten. Nach einer kurzen Einschulung kann es jeder Lernende auch als Nachschlagewerk nutzen.

Die einzelnen Korpora des **DWDS** beinhalten nur Texte aus Deutschland. Ein Teil von ihnen wird daher zum Bestandteil eines geplanten plurizentrischen Korpus mit einzelnen Varietäten des Deutschen. Dazu: DWDS (2014) unter Ressourcen → Korpora → 1. Referenzkorpora, Absatz C4-Korpus.

4.2.1.3 Korpus-C4

http://www.korpus-c4.org/
Dieses Korpus – eigentlich ein System von 4 Korpora – soll die deutsche Sprache des 20. Jahrhunderts plurizentrisch abdecken. Auf der Homepage des Projekts erfährt man Näheres unter den einzelnen Ikonen INFORMATIONEN TEILPROJEKTE STRUKTUR:

INFORMATIONEN	Zusammensetzung des Korpus (s. auch hier unter Struktur) und das Abfragesystem: Über die Zusammensetzung und das Abfragesystem wird wörtlich geschrieben (2014): **„Zusammensetzung** Am Korpus C4 beteiligt sind das >> Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS) aus Berlin, das >> Austrian Academy Corpus (AAC) aus Wien, das >> Korpus Südtirol aus Bozen und das >> SCHWEIZER TEXT KORPUS (CHTK) aus Basel. Das Projekt verfolgt das Ziel, die deutsche Standardsprache des 20. Jahrhunderts möglichst ausgewogen zu erfassen und online zugänglich zu machen. Momentan setzt sich das Korpus aus 20 Mio. Textwörtern des DWDS, 4.1 Mio. Textwörtern des AAC, 1.7 Mio. Textwörtern des Korpus Südtirol und 20 Mio. Textwörtern des CHTK zusammen. **Abfragesystem** Eine technische Besonderheit des Korpus C4 ist die verteilte Abfrage. Jedes Teilprojekt stellt seine Daten auf einem eigenen Server zur Verfügung, und erst zum Zeitpunkt der Abfrage werden die Resultate von den einzelnen Servern abgeholt und zusammengesetzt. Die Zusammenführung der Teilkorpora zu einem gemeinsamen ganzen Korpus C4 ist also nur virtuell. Technisch nutzt das Korpus dazu vor allem Funktionen der linguistischen Suchmaschine DDC, die vom Berliner Teilprojekt DWDS entwickelt wurde.“
TEILPROJEKTE	Beschreibung der einzelnen „Landeskorpora“
STRUKTUR	Aktuelle detaillierte Struktur der einzelnen Korpora in den einzelnen Quartalen des 20. Jahrhunderts

Ausgewogen ist im Moment nur der Teil des DWDS. Die anderen Korpora werden erst aufgebaut.
Die Ikone BEISPIELE zeigt die Abfragemöglichkeiten, die sich zwar im Wesentlichen mit den Abfragen an das DWDS decken (vgl. CHTK 2008-2014 → Beispiele), die Syntax der Abfragen weicht jedoch teilweise ab.

Abfragemöglichkeiten im Korpus-C4
Bei der Suche wird automatisch die Groß-/Kleinschreibung beachtet. Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis (Beschreibung)
Gärten	*Gärten* (nur diese Wortform)
$Lemma=Garten	*Garten, Gartens, Gärten …* (alle flektierten Formen von *Garten*)
Garten*	*Gartenkonzert*, *Gartengesellschaft*, *Gartenarbeit …*
***garten**	*Obstgarten*, *Versuchsgarten*, *Kindergarten …*
"schöne·Gärten"	*schöne Gärten* (Verbindung in exakt dieser Form)
"$Lemma=schön·$Lemma=Garten"	*schöner Garten*, schöne Gärten (Verbindung jeglicher Form von *schön* mit jeglicher Form von *Garten*)
"$Lemma=schön·#3·$Lemma=Garten"	*um den schönen Weg zwischen den Gärten (…) zu genießen* (Lemma *schön* und Lemma *Garten*, max. Abstand 3 Wörter)
wild·&&·Garten	*…der ungepflegte Garten wucherte wild und bunt …* (*wild* und *Garten* in einem Satz)
Haus·||·Garten	*freie Bewegung in Garten*, *Haus und um's Haus gestattet* (Sätze mit *Haus* oder *Garten* oder mit beiden Wörtern)
Garten·&&·!$Lemma=schön	*Garten* (Sätze mit *Garten*, jedoch ohne jegliche Form von *schön*)

Tab. 4: Abfragen Korpus C4

Eingabe der Abfrage
Zum Abfragefenster gelangt man über
Korpus → >C4-Suche beim SCHWEIZER TEXT KORPUS, Basel
Vor der Abfrage sind keine Einstellungen notwendig.

Beispiel: Suche nach der Wortform *Hit*

- **Einstellungen:** keine erforderlich
- **Abfrage** (Suchfeldeingabe): **Hit**

Ergebnis:

1975	… Der Partnertausch ist tot — es lebe der Lebenstausch! Der neueste	**Hit**	unter emanzipationshungrigen Frauen ist der Identitätswechs
1976	… Ihr Bruder Markus stellt den Kassettenrecorder an und lässt einen	**Hit**	laufen. Doris schreit Markus an, er solle sofort mit dem Lärm …
1976	… auf sie, es gibt Streit. 1 Kann jemand von euch die Fremdwörter	**Hit**	und Kassettenrecorder erklären? Was heisst aufhören im …
1976	… fertig. Sie steht auf, stellt den Kassettenrecorder an, lässt den	**Hit**	laufen und pfeift vergnügt dazu. Markus sagt: « Du bist aber …
1979	… Der Kasperli sang bei seinem ersten Auftritt seinen jahrelangen	**Hit**	« Trullalla, trullalla, de Chaschperli isch wieder da », worauf im
1986	… eindringen, anprangern, all die clever, cool, Boß, Dreß, Fan,	**Hit**	, Job, Quiz, City, Spray, Team, Trend, Boom, Manager; was uns
1989	… nach einem Monat buchstäblich plattgewalzt. Die Lattenskis, ein	**Hit**	der Bewegungsspiele, waren dieser so grossen Belastung eber
1991	… Zeitungspapier am Laufmeter, auch das ist spannend; der absolute	**Hit**	waren aber Holzrugel, ein Wort das mittlerweile auch in Amer
1993	… die Vereinsmänner immer wichtigere Preise auspackten: Der grosse	**Hit**	war ein Schwarzweiss-Fernseher. Und eine …

Abb. 40: Konkordanzen zur Abfrage: Wortform *Hit* (CHTK)

Die abgerufenen Konkordanzen lassen sich filtern und sortieren: über die Ikonen Filter Sortieren Export Optionen Beispiele werden einfache und übersichtliche Formulare abgerufen, die man anwenden kann.

Auch das Export-Format (TXT) ist sehr gut und einfach lesbar (auch wenn es auf den ersten Blick nicht so aussieht): Die Daten werden aus einem Notizblock (TXT) einfach in eine Excel-Tabelle kopiert (TXT: Strg+A → Strg+C → EXCEL: Strg+V). In Excel werden die folgenden Angaben ausgewiesen: die gewählten Metainformationen (Spalte A), der Kontext links vom KWIC (Spalte B), der Kontext rechts vom KWIC (Spalte C), wie in der Abb. 40 dargestellt wird. So lassen sich die Konkordanzen sehr einfach sortieren.

Das Korpus C4 würde sich sehr gut für die plurizentrische Arbeit eignen (zugänglich und recherchierbar ist im Moment das Schweizer Textkorpus, aus dem die Konkordanzen in der Abb. 40 stammen). Leider sind die einzelnen Teile noch nicht ausgewogen und wegen der Größe der Teile ist man auch weit von der Repräsentativität des sprachlichen Materials entfernt. Für die plurizentrische Arbeit wird jedoch das DeReKo empfohlen.
Mangelhaft bzw. gar nicht vorhanden sind im C4 die statistischen Funktionen. Auf der anderen Seite ist das Korpus sehr einfach und intuitiv zu bedienen; es eignet sich dadurch als ein guter Einstieg in die Korpuslinguistik.

4.2.1.4 Das Deutsche Referenzkorpus (DeReKo)

https://cosmas2.ids-mannheim.de/cosmas2-web/
Das Deutsche Referenzkorpus ist auch unter dem Namen **IDS-Korpus** oder **Cosmas II** bekannt. Die Bezeichnungen sind jedoch nicht ganz richtig: das Institut für Deutsche Sprache in Mannheim entwickelt und verwaltet mehrere Korpora (z.B. auch die DGD), Cosmas II hingegen ist ein vom IDS entwickelter Korpusmanager (Corpus Search, Management and Analysis System, Version II). Unter diesem Manager laufen alle IDS-Korpora der geschriebenen Sprache.

Das DeReKo ist momentan das größte Korpus für Deutsch weltweit, es enthält über 6 Milliarden Textwörter (Stand 2013) und wird sukzessive und regelmäßig erweitert. Die Bedienung aller IDS-Korpora ist ziemlich komplex. Für jede/-n, der/die sich mit der deutschen Sprache befasst, ist es dennoch, ratsam zumindest die Basisfunktionen von Cosmas II (Korpusmanager) sicher zu beherrschen. Es gibt leicht zugängliche Quellen, die die Bedienung sehr gut beschreiben: als Einstieg empfiehlt sich die Anleitung von Stephen Berman (2013). Weiterhin verfügt Cosmas II über ein sehr gutes Hilfe-Portal (Cosmas II 2012). Das Wesentliche (kombiniert aus beiden Quellen und ergänzt) wird hier beschrieben.

Der Zugang zum Korpus der geschriebenen Sprache erfolgt entweder direkt über die Internet-Schnittstelle (http://www.ids-mannheim.de/cosmas2/web-app/) oder über eine Windows-Applikation, die auf die Festplatte heruntergeladen werden muss.

Installieren der Cosmas II_{win} Suchmaschine

Ins Google- oder Adressen-Suchfeld **http://www1.ids-mannheim.de/start/**	eingeben
Forschung → laufende Projekte → Cosmas II → Cosmas II_{win}	anklicken.
Detaillierte Informationen	
Applikation herunterladen und installieren (…) .	anklicken.

Beim Herunterladen und Installieren der Anleitung folgen. Es ist sehr einfach!

Zugang zur Cosmas II_{web}-Schnittstelle

Ins Google- oder Adressen-Suchfeld **http://www.ids-mannheim.de/cosmas2/**	eingeben
Cosmas II_{web}	anklicken.
Aktuelle Version:	
Applikation starten.	anklicken

Für die Recherche sowohl über die Suchmaschine als auch über das Web benötigt man eine Registrierung.

Die Funktionen von CosmasII$_{win}$ und CosmasII$_{web}$ sind fast ident. CosmasII$_{win}$ ermöglicht die sog. „graphische Suche“: das gewünschte Kästchen (links) auf die graue Fläche ziehen, durch einen Doppelklick öffnen und ins Suchfeld den Suchbegriff eingeben. Des Weiteren hat es einige zusätzliche Funktionen, die CosmasII$_{web}$ nicht hat. Für den ersten Einstieg wird CosmasII$_{web}$ empfohlen, die folgende Beschreibung bezieht sich terminologisch auch auf die Web-Applikation.

Abfragen
Die sprachlichen Elemente, die sich mit Cosmas II abfragen lassen, sind übersichtlich unter: Online Hilfe → Suchanfrage → Zeilenangabe → Beispiele aufgeführt. Die Syntax der häufigsten Abfragen ist hier mit Erklärungen ab Seite 60 zu finden.
Im Folgenden werden einige Recherchen Schritt für Schritt beschrieben, somit wird die (oft mühsame) Suche im umfangreichen Hilfe-Portal hinfällig.

4.2.1.4.1 Recherche in allen Archiven

Einstieg: Anmeldung → Login → Recherche
Der Weg zur Abfrage geht über die folgenden Schritte (A.→ B.→ C.→ D.):

A. **Auswahl des Archivs**
B. **Auswahl des Korpus**
C. **Einstellung der Optionen**
D. **Suchanfrage**

A. Auswahl des Archivs

DeReKo Archive:

W - Archiv der geschriebenen Sprache Enthält alle zugänglichen Texte (über 4 Mrd. [1)] Wörter). **Empfohlen** für übliche Recherchen.
W-ÜBRIG - Archiv der aussortierten geschriebenen Korpora
TAGGED-C - Archiv morphosyntakt. annotierter Korpora (CONNEXOR) Enthält etwa ¼ der Texte aus dem Archiv W (über 1 Mrd. [1)] Wörter).
TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)[2] Enthält etwa ¼ der Texte aus dem Archiv W (über 1 Mrd. [1)] Wörter). **Empfohlen** für Recherchen mit morphosyntaktischen Annotationen.
TAGGED-M - Archiv morphosyntakt. annotierter Korpora (MECOLB) [3] Enthält ca. 20 Mio. [1)] Wörter aus dem Archiv W.
HIST - Archiv der historischen Korpora Texte bis 1962
GFDS - Kartei der Gesellschaft für deutsche Sprache
WK-PH - Archiv der phasengegliederten Wendekorpora

Anmerkungen:
1) Die Angaben über die Größe der Archive beziehen sich nur auf öffentlich zugängliche Texte.
2) Das Archiv TAGGED-T wird deswegen empfohlen, da es mit demselben Tagset annotiert wird als andere hier erwähnte Korpora (DWDS, InterCorp).
3) Das Archiv TAGGED-M ermöglicht auch Abfragen nach morphologischen Kategorien, die sonst nicht abgefragt werden können (Kasus, Genus). Es beinhaltet nur bundesdeutsche Texte und ist sehr klein.

B. Auswahl des Korpus

Empfehlenswert ist alle vordefinierten/ alle öffentlichen Korpora auszuwählen**.**
Beliebige Subkorpora kann man über die Ikone Korpusverwaltung erstellen.

C. Einstellung der Optionen

Vor der Sucheingabe soll zuerst auch der Suchmodus eingestellt werden (unter Optionen). Bitte nicht vergessen, nach der jeweiligen Einstellung die Ikone Übernehmen anzuklicken.
Empfehlung: um die Einstellung für die nächste „Sitzung" zu speichern, muss nach Abschluss der Recherchen die Webseite über Logout verlassen werden. Nach dem einfachen Schließen der Web-Seite werden die Einstellungen auf die Standardwerte zurückgesetzt.

Die meisten Optionen sind selbsterklärend. Kommentiert werden hier nur die erfahrungsgemäß problematischen Punkte.

Suche	
Suchmodalitäten	
Groß- / Kleinschreibung beachten für 1. Zeichen	aus-/einschalten
Groß- / Kleinschreibung beachten für andere Zeichen	aus-/einschalten
Diakritische Zeichen beachten	aus-/einschalten
Expansionslisten (Liste der Formen)	
Expansionslisten anzeigen	aus-/einschalten
mit Häufigkeiten (langsam)	alternativ aus-/abwählen
ohne Häufigkeiten (mittel)	alternativ aus-/abwählen
ohne Korpusvalidierung (schnell)	alternativ aus-/abwählen
Sortierung[1)]	Auswahl
Zusammenfassung ab	Zahl eingeben
Begrenzung der Ergebnismenge (Falls zu viele zu erwarten sind.)	
alle Treffer	alternativ aus-/abwählen
Zufallsauswahl	alternativ aus-/abwählen
durch Zufallsauswahl eingrenzen auf:	Zahl eingeben
Zufallsgenerierung: fest[2)] variabel	alternativ aus-/abwählen
	Übernehmen anklicken

Anmerkungen:
1) Nach jeder Präsentation der Ergebnisse setzt sich die Sortierung auf alphabetisch aufsteigend zurück.
2) Auswahlmöglichkeit fest bedeutet: bei der nächsten Abfrage desselben Phänomens bekommt man dieselben Konkordanzen

Lemmatisierung	
Lemmatisierung	
Flexionsformen	immer eingeschaltet
Komposita	aus-/einschalten
Sonstige Wortbildungsformen[1)]	aus-/einschalten
Spezialfälle[2)]	aus-/einschalten
	Übernehmen anklicken

Anmerkungen:
1) Einschalten, wenn z.B. nach Affixen gesucht wird.
2) Wörter mit Bindestrich, Nummern etc.

Korpuspräsentation	
Korpuspräsentation	
Korpusansicht	Auswahl
Sekundäre Sortierung:	Auswahl
Prozentuelle Verteilung	aus-/einschalten
Darstellung	aus-/einschalten
	Übernehmen anklicken

Ergebnispräsentation	
Ergebnispräsentation	
Ergebnisansicht:	Auswahl
Sekundäre Sortierung der Ergebnisse:	Auswahl
Berechnung des Häufigkeitsmaßes (Frequenzen und Distribution)	
Maß für die Häufigkeit berechnen und anzeigen	aus-/einschalten
Relative Häufigkeiten	alternativ aus-/abwählen
in Prozent pro Million Worte	alternativ aus-/abwählen
Differenzenkoeffizient	alternativ aus-/abwählen
Häufigkeitsklassen mit folgender Referenz:	alternativ aus-/abwählen
Einzelwort, Groß/Klein/Diakr. beachten Einzelwort, Groß/Klein/Diakr. ignorieren Lemma automatisch	alternativ aus-/abwählen
Darstellung: Häufigkeitsmaß grau darstellen	aus-/einschalten
	Übernehmen anklicken

KWIC/Volltext	
Treffermarkierung: Markierungsfarbe	Auswahl
KWIC-Anzeige: Volltext-Anzeige Angaben über die Präsentation der Konkordanzen und Textabschnitte	Eingaben und Auswahl
Seitengröße	Zahl eingeben
Alphabetische Sortierung[1)]	
1.Kriterium:	immer eingeschaltet/ Auswahl
2.Kriterium:	aus-/einschalten/ Auswahl
3.Kriterium:	aus-/einschalten/ Auswahl
Sonderzeichen ignorieren	aus-/einschalten
Zufällige Sortierung: fest variabel[2)]	alternativ aus-/abwählen
	Übernehmen anklicken

Anmerkungen:

1) Sortierung nach der unmittelbaren Umgebung des KWICs (bis zu 3 Positionen links oder rechts).
2) Auswahlmöglichkeit fest bedeutet: bei der nächsten Abfrage desselben Phänomens bekommt man dieselben Ergebnisse.

D. Suchanfrage

Die Eingabe der Abfrage erfolgt über den Button Suchanfrage. Beim Schreiben der Sucheingabe ins Suchfeld muss man immer den **Abstand zwischen den einzelnen Wörtern/ Zeichenketten und Sonderzeichen beachten!** Leerzeichen werden hier als ein Mittelpunkt (·) gekennzeichnet.

Beim Suchen nach Elementen mit Abstand (Wortkombinationen) immer die folgende Funktion überprüfen:

Weggelassener Verknüpfungsoperator bedeutet:	
⊙ Wortabstand /+w1 ○ logisches "ODER"	alternativ aus-/abwählen

Die Abfrage wird ins Suchfeld **Eingabe** geschrieben und durch Klicken auf Suchen aktiviert.

Eingabe: **irgendwas**
Suchen anklicken (Ergebnisse anklicken)

Die Ergebnisse sind Konkordanzzeilen, die dann sortiert und gespeichert werden können (Abb. 41).

Konkordanzen (10 Beispiele, Zufallsauswahl)**:**

```
A13        Einzelhaft gesetzt zu werden, weil irgendwas passiert war. Einmal sprangen alle
A13            einer Band, die niemandem mehr irgendwas beweisen und keinen zeitgeistigen
A13             sein.» Wo steht hier irgendwo irgendwas von Koran, Hadith oder Teilauszüge
A13                    auch nur, um sich vor irgendwas zu drücken, oder? Das Weltgeschehen
A13             hochgeladen, in dem ein Model irgendwas erzählt. Frei nach dem Motto:
A13                    des Weges noch im Tal irgendwas. Dafür sahen sie sich tief
A13               wollen die Jungen häufig in irgendwas belehren. Die Jugendlichen wissen
A13                  Liner hat zu keiner Zeit irgendwas des Geldes willen gemalt. Mehrmals
A13 Rechtschreibung jedenfalls. Heureusement, irgendwas mit Stunden, schlägt Bazyli vor.
A13            Manchmal merke ich mir einfach irgendwas nicht, und ich habe dann
```

Abb. 41 Konkordanzen zur Abfrage *irgendwas* (DeReKo, W-Archiv)

Abfragemöglichkeiten im Cosmas II

Die Möglichkeiten der Abfragen weichen von den üblichen Korpusabfragetypen nicht ab. Die Syntax der Abfragen mag jedoch am Anfang etwas ungewohnt wirken. Für eine bessere Orientierung werden die Abfragen hier in vier Gruppen mit steigender Komplexität aufgeteilt. In der rechten Spalte werden typische Ergebnisse und ein Kommentar angeführt.

Suche nach:

1) einzelnen Wörtern bzw. Zeichenketten, die Bestandteil eines Wortes sind;
2) Wortkombinationen;
3) Kombinationen mit Abstand;
4) morphosyntaktischen Kategorien (in Kombination mit Abstand).

Abfragen: Wörter/ ununterbrochene Zeichenketten

Eingabe	Ergebnis, Kommentar
Arzt	Nur die Wortform *Arzt*, jedoch nach Einstellung der Suchmodalitäten (Optionen – Suchmodalitäten). Falls alles abgewählt, dann auch *Ärzt, ARZT* oder *ARzt* etc.
&Arzt	*Arzt* lemmatisiert (alle flektierten Formen von *Arzt*): *Arzt, Arztes, Ärzte …*
Arzt*	Alle Wörter, die mit *Arzt-* beginnen: *Arzt, Arztpraxen, Arztpraxis, Arztbesuch, Arzthelferin* sind die häufigsten im DeReKo, W-Archiv.
***arzt**	Alle Wörter, die auf *-arzt* enden: *Notarzt, Chefarzt, Hausarzt, Zahnarzt, Tierarzt, Facharzt, Oberarzt, Kinderarzt, Augenarzt* sind die häufigsten im DeReKo, W-Archiv.
++art	Alle Wörter, die auf *-art* enden und maximal aus 5 Buchstaben bestehen: *Art, hart, Wart, Chart, smart* sind die häufigsten im DeReKo, W-Archiv.
?art	Alle Wörter, die auf *-art* enden und genau aus 4 Buchstaben bestehen: *hart, Part, Bart, zart…* sind die häufigsten im DeReKo, W-Archiv.
&-schaft	Alle Wörter mit dem Suffix *-schaft*: *Mannschaft, Gesellschaft, Wirtschaft, Staatsanwaltschaft, Meisterschaft…* sind die häufigsten im DeReKo, W-Archiv.
&wider-	Alle Wörter mit dem Präfix *wider-*: *Widerstand, widersprachen, widerlegt, Widerlegung, Widerhaken* sind die häufigsten im DeReKo, W-Archiv.

Tab. 5: Abfragen COSMAS II (Wörter)

Erklärungen und Bemerkungen:

Symbol	Bedeutung
&	„Lemmatisator“: Sucht alle Formen des Grundwortes bzw. alle Wörter mit dem eingegebenen Affix[16]. Kein Leerzeichen zw. & und der Grundform, bzw. dem Affix.
*	„Platzhalter“: öffnet die Grenze der Zeichenkette um eine beliebige Anzahl an Zeichen
+	„Platzhalter“: öffnet die Grenze der Zeichenkette um ein oder kein Zeichen.
?	„Platzhalter“: öffnet die Grenze der Zeichenkette um genau ein beliebiges Zeichen.

Abfragen: Wortkombinationen (ununterbrochene Zeichenketten)

Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis, Kommentar
bester·Arzt	Genaue Verbindung von den Wörtern: *bester Arzt*
&gut·&Arzt	Verbindung *guter Arzt* in unterschiedlichen Formen: *guter Arzt, gute Ärzte, besten Ärzte* sind die häufigsten Formen im DeReKo, W-Archiv
d++·&gut·&Arzt	Verbindung *guter Arzt* in unterschiedlichen Formen mit bestimmtem Artikel (oder einem anderen Wort, das auf *d* beginnt und max. 3 Buchstaben hat): *die besten Ärzte, der beste Arzt, den besten Ärzten* sind die häufigsten Formen im DeReKo, W-Archiv.
Kanzlerin·oder·Merkel	Alle Texte, in denen die Wörter *Kanzlerin* oder *Merkel* vorkommen.
Merkel·„oder“	Genaue Verbindung von *Merkel* und *oder*: *… **Merkel oder** Außenminister …*; ***Merkel oder** Bundespräsident*; *… **Merkel oder** Edmund Stoiber …*
Kanzlerin·nicht·Merkel+	Alle Texte, in denen *Kanzlerin*, aber nicht *Merkel/ Merkels* vorkommen.
Merkel·und·Obama	Alle Texte, in denen die Wörter *Merkel* und *Obama* vorkommen.

Tab. 6: Abfragen COSMAS II (Wortkombinationen)

Erklärungen und Bemerkungen:

Symbol	Bedeutung
oder, nicht, und	Diese Wörter (und nur diese) sind sogenannte „logische Operatoren“, deswegen müssen sie in Anführungszeichen gesetzt werden, wenn wir sie abrufen wollen (mehr dazu siehe Cosmas II → Hilfe → Logische Operatoren)

Abfragen: Kombinationen mit Abstand – allgemein

Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis, Kommentar
Kanzlerin·/w1·Merkel	Zwei nebeneinanderstehende Wörter: *Kanzlerin Merkel, Merkel Kanzlerin* (*Kanzlerin* und *Merkel* im Abstand von einem Wort, beliebige Reihenfolge)
Kanzlerin·/+w1·Merkel	Zwei nebeneinanderstehende Wörter: *Kanzlerin Merkel* in dieser Reihenfolge
Kanzlerin·/-w1·Merkel	Zwei nebeneinanderstehende Wörter: *Kanzlerin Merkel* in umgekehrter Reihenfolge, also nur *Merkel Kanzlerin*
Merkel·/s0·Obama	Sätze, in denen die Wörter *Obama* und *Merkel* (in beliebiger Reihenfolge) vorkommen.
Kanzlerin·%Merkel	Sätze, in denen das Wort *Kanzlerin* vorkommt, jedoch nicht das Wort *Merkel*.

Tab. 7: Abfragen COSMAS II (Kombinationen mit Abstand)

[16] Bei der Affixsuche muss die Option „sonstige Wortformen“ eingeschaltet werden (Optionen → Lemmatisierung → Sonstige Wortbildungsformen). Die Liste mit suchbaren Affixen ist auch hier im Kap. 8.1.

Erklärungen und Bemerkungen

Symbol	Bedeutung
/w2	Dieses Symbol bedeutet Wortabstand (**w** für „Wort“) zwischen zwei Elementen. Die Ziffer (hier 2) bedeutet den Abstand von zwei Wörtern, d.h. zwischen den abgefragten Wörtern liegt noch ein anderes. Die Reihenfolge ist beliebig. Die Reihenfolge wird durch die Plus- und Minuszeichen bestimmt. Das Pluszeichen (+) bedeutet „diese Reihenfolge“, das Minuszeichen (-) bedeutet „umgekehrte Reihenfolge“ (siehe Ergebnis Kanzlerin·/+w1·Merkel und Kanzlerin·/-w1·Merkel)
/s0	Das Zeichen **s** bedeutet „Satz“, die Ziffer 0 denselben Satz. Analog zu Wort (w) können Elemente in nacheinander folgenden Sätzen gesucht werden (/s1 bedeutet in 2 nacheinander folgenden Sätzen, /s2 im Abstand von maximal 2 Sätzen etc.)
/p0	Das Zeichen **p** bedeutet „Absatz“. Analog zu den Abstandsoperatoren w und s können Elemente auch im Abstand von Absätzen gesucht werden.

Abfragen: Kombinationen mit Abstand – strukturelle Fragen innerhalb eines Satzes

Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis, Kommentar
&erwidern·/w0·,	Suche nach dem Wort *erwidern* (in unterschiedlichen Konjugationsformen), nach dem ein Beistrich folgt: ***erwidert,*** *aber …*; ***erwiderte,*** *dass …*
&fordern·/s0·auf	Suche nach Verben mit (trennbarem) Verbzusatz (traditionell „trennbares Präfix“ genannt). Diese Abfrage findet jedoch auch Sätze, in denen beide Elemente vorkommen, bilden gemeinsam jedoch nicht das Lexem auffordern. *(Sie **fordert** zwar einen Rechtsanspruch zur Rückkehr **auf** eine Vollzeitstelle,…).* Deswegen ist es sinnvoller die folgende Abfrage zu wählen:
&fordern /s0 auf #IN(R)·<s>	Suche nach Verben mit (trennbarem) Verbzusatz (traditionell „trennbares Präfix“ genannt), wobei der Verbzusatz am Ende des Satzes steht.
Wo+++·/s0·\?	Fragesätze, die mit *Wo, Wozu, Wofür …* beginnen.
Wie+++·/s0·!	Exklamative Sätze beginnend mit *Wie, Wieso …*
Wer·/s0·.	Aussagesätze, die mit *Wer* beginnen.
Wer·#IN(L)·<s>·/s0·gedreht	*Wer* am Anfang des Satzes, in demselben Satz auch das Partizip Perfekt von *drehen*: ***Wer** hat an der Uhr **gedreht**? Wer hat hier denn seine Runden gedreht? Wer wissen und vor allem sehen möchte, wo berühmte Filme gedreht wurden,*
Wer·#IN(L)·<s>·/s0·gedreht·#IN(R) <s>	Suche nach *Wer* am Anfang und *gedreht* am Ende desselben Satzes: ***Wer** hat daran **gedreht**? **We**r hat an der Uhr **gedreht** …*
&Einbahn·#IN <ü>	Suche nach allen Formen des Wortes *Einbahn* in allen Überschriften.

Tab. 8: Abfragen COSMAS II (Kombinationen im Satz)

Erklärungen und Bemerkungen

Symbol	Bedeutung
#IN(L) <s>	Dieses Symbol hinter einem Abfrageelement (hier Wer) bedeutet Anfang eines Satzes (L=“links“).
#IN(R) <s>	Dieses Symbol hinter einem Abfrageelement (hier *gedreht*) bedeutet Ende eines Satzes (R=“rechts“).
<s>	Dieses Symbol bedeutet Satz. Es können auch andere Textteile abgerufen werden:
<ü>	Alle Überschriften (zu Unterkategorien siehe „Suchanfragen in Überschriften“ in Cosmas II Online-Hilfe).

Mehr zur Abfragesyntax im COSMAS II unter: Cosmas II_{web} → Online-Hilfe → Suchanfrage → Zeileneingabe → Syntax.

4.2.1.4.2 Recherche im morphosyntaktisch annotierten Teil

Viele deutsche Wörter oder Verbindungen sind mehrdeutig. Um die Mehrdeutigkeit (Polysemie oder Homonymie) zu verringern, empfiehlt es sich, ein Korpus mit Tagging zu verwenden. Tagging ist ein Prozess, in dem jedes Wort mit einer Zusatzinformation versehen („annotiert") wird. Diese Information bezieht sich auf die morphosyntaktische Charakteristik des Wortes. Im Prinzip handelt es sich um morphologische, teilweise auch syntaktische Kategorien, die aus der Form und Position des Wortes im Satz abgeleitet werden können (Wortart, Numerus, Tempus etc.).

Unter Cosmas II gibt es drei morphologisch annotierte (getaggte) Korpora: TAGGED-C, TAGGED-T und TAGGED-M.
Zur Abfrage nach Tags (morphosyntaktischen Kategorien) verwendet man in beiden Versionen von Cosmas II (Cosmas II_{web} und Cosmas II_{win}) die sog. „graphische Suche" - man wählt aus dem Angebot (Pop-up-Fenster) des morphologischen Assistenten die gewünschte Kategorie. (Im Cosmas II_{win} ist die graphische Suche auch im nicht-annotierten Teil möglich.)

Das IDS empfiehlt „die annotierten Korpora der neueren Archive TAGGED-C oder TAGGED-T (Stand: 2010 bzw. 2011) zu verwenden." (IDS online[17]). Dennoch kann auch das Archiv TAGGED-M interessant sein – es ist zwar im Vergleich sehr klein (ca. 30 Mio. laufende Wortformen), bietet aber andere Funktionen, die die größeren Archive nicht ermöglichen (siehe Thematische Auflistung der morphosyntaktischen Tags im Kap. 8.2 - in der Tabelle sind die abfragbaren Kategorien im MECOLB (Archiv TAGGED-M) und Treetagger (Archiv TAGGED-T) nebeneinander gestellt, in der letzten Spalte die Kürzel im InterCorp).

Abfragen
Zuerst muss das entsprechende Archiv und Korpus ausgewählt werden:

Auswahl des Archivs und Korpus

Archiv – TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)	auswählen
Korpus - TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T	auswählen

In getaggten Archiven kann man dieselben Abfragen stellen wie im nicht-getaggtem Archiv. Beim Schreiben der Sucheingabe ins Suchfeld muss man immer den **Abstand zwischen den einzelnen Wörtern/ Zeichenketten und Sonderzeichen beachten!** Leerzeichen werden hier als ein Mittelpunkt (·) gekennzeichnet.

Beim Suchen nach Elementen mit Abstand (Wortkombinationen) immer die folgende Funktion überprüfen:

Weggelassener Verknüpfungsoperator bedeutet:	
⊙ Wortabstand /+w1 ○ logisches "ODER"	alternativ aus-/abwählen

Für das kommende Beispiel muss Wortabstand /+w1 ausgewählt werden.

[17] Bemerkungen zum Umgang mit morphosyntaktisch annotierten Korpora (Cosmas II → Textorganisation → Annotationen → Bemerkungen)

Beispiel: Gesucht wird nach dem Wort *um* in der Rolle eines Verbzusatzes (***um****denken*, ***um****kreisen*…), der abgetrennt ist.

Zuerst wird der morphosyntaktische Assistent aktiviert:

MORPH-Assistent	anklicken

Jetzt erscheint ein Pop-up-Fenster, in dem das gewünschte Phänomen ausgewählt werden kann.
Die Rolle „abgetrennter Verbzusatz“, nach dem jetzt gesucht wird, ist unter Partikel, abgetrennter Verbzusatz zu finden. Diese muss man „übernehmen“.

Partikel → abgetrennter Verbzusatz	Übernehmen anklicken

Wenn man jetzt auf Suchen klicken würde, bekäme man als Ergebnis alle abgetrennten Verbzusätze (*…fest, …ein, …um* etc.). Gesucht wird jedoch lediglich nach … *um*, deswegen muss die Abfrage um ein Argument ergänzt (kombiniert) werden.

Kombinationen:
Wenn ein konkretes Wort in einer bestimmten morphosyntaktischen Rolle (als eine „morphologische Kategorie“) gesucht wird (hier: *um* als abgetrennter Verbzusatz), muss die Abfrage als „Null-Abstand“ formuliert werden.
Nachdem der morphologische Assistent das entsprechende morphosyntaktische Zeichen (Tag) ins Suchfeld eingegeben hat (hier: MORPH(PTK vz)), müssen noch der „Null-Abstand“ und das gewünschte Wort oder Element (hier: *um*) eingegeben werden:

/w0	(Abstand „0“) eingeben
um	(Wort *um*) eingeben

Die Eingabe sieht wie folgt aus (Abstand beachten!):

Eingabe: **MORPH(PTK vz)·/w0·um**	
	Suchen anklicken

Bitte bedenken Sie, dass die morphologische Annotation automatisch durchgeführt wurde und die Ergebnisse eine große Fehlerquote aufweisen. Bedenken Sie während der Aussortierung falscher Belege, wie lange eine manuelle Recherche in Büchern, Zeitungen und Zeitschriften dauern würde.

Abfragen mit Abstand: Kombinationen mit Tag (über morphologischen Assistenten) (Auswahl)
Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis, Kommentar
MORPH(N ne)·/w0·König	*König* als Eigenname: *Kardinal Franz* ***König***; *Trainer Christian* ***König***
MORPH(N nn)·/w0·König	*König* als Gattungsname (im morphosyntaktischen Assistent als „normales Nominum“ bezeichnet): *Simba –* ***König*** *der Löwen*; *der Heilige* ***König*** *…*
MORPH(AP pr)·/+w2·König	*König* im Abstand von max. 2 Wörtern nach einer beliebigen Präposition: … ***durch König*** *Karl I. …*; … ***durch*** *den* ***König*** *Herodes …*; ***Laut*** *SP-Bürgermeister König …*

(Fortsetzung auf nächster Seite)

Eingabe	Ergebnis, Kommentar
MORPH(AP·pr) /+w2·(König·/w0·MORPH(N·nn))	
	König als Gattungsname im Abstand max. 2 Wörter davor eine beliebiger Präposition: … ***auf*** *den* ***König*** *…, … **mit** einem **König** …*
MORPH(PTK·vz) /w0·um	*um* als Verbzusatz
MORPH(VRB·fin·a)·MORPH(VRB·inf·v) MORPH(VRB·inf·m) #IN(R) <s>	
	Verbalkomplex (dreistellig) am Ende eines Satzes: *da diese bereits vor vier Wochen hätte erfolgen müssen.*

Tab. 9: Abfragen COSMAS II (Kombinationen mit Tag)

4.2.1.4.3 Ansichten

Diese Funktion ist wichtig für die Entdeckung der Unterschiede im Vorkommen der KWICs. Die Ansicht kann nach unterschiedlichen Kriterien ausgewählt werden, die die Distribution des KWICs auf der stilistischen (Themen/Textsorten), regionalen (Länderansicht) oder historischen (Jahresansicht) Ebene beleuchten. Über Cosmas II_{web} kann man unter folgenden Ansichten wählen: Quellenansicht, Korpusansicht, Dokumentansicht, Ansicht vor/ seit (Zeitpunkt), Jahrzehntansicht, Jahresansicht, Monatsansicht, Tagesansicht, Länderansicht, Textsortenansicht, Themenansicht. Im Cosmas II_{win} gibt es dazu noch die Ansicht nach Wort-Types (Statistik einzelner Formen z.B. bei einer Lemmaabfrage). Unter Ergebnisse kann zwischen den einzelnen Ansichten durch Klicken auf die entsprechende Ikone gewechselt werden.

Beispiel: Gesucht wird nach den Lexemen *Gehsteig* (Abb. 42) und *Gummibärli* (Abb. 43) mit dem Ziel zu recherchieren, ob ihr Vorkommen regional bedingt ist.

Länderansicht:

Zuerst soll wieder das entsprechende Archiv und Korpus ausgewählt werden, erst dann werden die gewünschten Optionen eingestellt. Die Einstellung der Optionen ist für die Präsentation der Ergebnisse nach einzelnen Kriterien (hier Länderansicht) elementar.

Auswahl des Archivs und des Korpus

Archiv – W - Archiv der geschriebenen Sprache	auswählen
Korpus - N-öffentlich - alle öffentlichen Korpora	auswählen

Einstellung:

Optionen → Korpuspräsentation

Korpuspräsentation	
Korpusansicht	**Länderansicht**
Sekundäre Sortierung:	**Texte (absteigend)**
Prozentuelle Verteilung der Texte	einschalten
Darstellung	einschalten
	Übernehmen anklicken

Optionen → Ergebnispräsentation

Ergebnispräsentation	
Ergebnisansicht:	**Länderansicht**
Sekundäre Sortierung der Ergebnisse:	**Texte (absteigend)**
Berechnung des Häufigkeitsmaßes (Frequenzen und Distribution)	
Maß für die Häufigkeit berechnen und anzeigen	einschalten
Relative Häufigkeiten	auswählen
in Prozent	
pro Million Worte	auswählen
Differenzenkoeffizient	
Häufigkeitsklassen mit folgender Referenz:	
Einzelwort, Groß/Klein/Diakr. beachten	auswählen
Einzelwort, Groß/Klein/Diakr. ignorieren	auswählen
Lemma	
automatisch	
Darstellung: Häufigkeitsmaß grau darstellen	
	Übernehmen anklicken

Eingabe der Abfrage:

Ins Suchfeld *Gehsteig* (lemmatisiert) eingeben:
Die Eingabe sieht wie folgt aus (Abstand beachten!):

&Gehsteig
Suchen anklicken

Nach dem eine Listen mit Formen (*Gehsteige, Gehsteigs* etc.) erscheint:.

Ergebnisse anklicken

Als Ergebnis bekommt man eine Tabelle mit Länderansicht, sortiert nach relativer Häufigkeit:

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	10.633	**15.61 pMW**	7.768	1991	2013	A
⊞	2.349	**0.75 pMW**	2.079	1949	2013	D
⊞	171	**0.37 pMW**	160	1996	2013	CH
	13.153	**3.07 pMW**	**10.007**	**1949**	**2013**	**3 Länder**

Abb. 42: *Gehsteig* (lemmatisiert), Länderansicht (DeReKo, W-Archiv)

Beim Klicken auf ⊞ eröffnen sich die Konkordanzzeilen (Belege).
Es ist ratsam, jeweils die relativen Häufigkeiten zu beobachten. Das Beispiel in der Abb. 42 zeigt: die meisten Treffer (2. Spalte = absolute Anzahl der „Treffer"/Belege im Korpus) kommen in österreichischen Texten vor und haben dort auch die höchste relative Häufigkeit (15.61 Vorkommnisse pro eine Million Worte).

Das Beispiel in Abb. 43 (gesucht wurde nach *Gummibärli*) könnte jedoch bei der simplen Betrachtung der absoluten Zahlen (Treffer) den Beobachter in die Irre führen:

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	85	**0.1834 pMW**	76	1996	2013	CH
⊞	93	**0.1365 pMW**	82	1992	2012	A
⊞	6	**0.0019 pMW**	5	2008	2011	D
	184	**0.0430 pMW**	**163**	**1992**	**2013**	**3 Länder**

Abb. 43: *Gummibärli*, Länderansicht (DeReKo, W-Archiv)

Die höchste absolute Trefferanzahl ist zwar wieder in österreichischen Texten zu verzeichnen, nach relativer Häufigkeit sieht man aber, dass das Wort *Gummibärli* (nicht überraschend) am häufigsten in schweizerischen Texten vorkommt.

4.2.1.4.4 Kookkurrenzanalyse

Zur Entdeckung von häufigsten Verbindungen und sprachlichen Chunks ist die Berechnung der Verbindungen einzelner Lexeme wichtig. Im DeReKo heißt diese Berechnung Kookkurrenzanalyse. Diese kann zu jedem Rechercheergebnis durchgeführt werden. Diese Korpusfunktion ist im (Fremd)Sprachenunterricht für die Erschließung der Redundanzfelder beim Leseverstehen (im Sinne von Westhoff (1987: 41) wichtig.

Beispiel: Gesucht wird nach dem Verb *aufsetzen* (ohne abgetrenntes *auf*) und seinen Kookkurrenzpartnern. In anderen Worten fragen wir das Korpus: „Was setzt man am häufigsten in deutschen Texten auf?“

Abfrage:
Auswahl des Archivs und des Korpus

Archiv – W - Archiv der geschriebenen Sprache	auswählen
Korpus - N-öffentlich - alle öffentlichen Korpora	auswählen

Einstellung der Optionen: Standardwerte

Eingabe: **&aufsetzen**
Suchen anklicken

Aus der Liste der Formen (*aufgesetzt*, *aufsetzen*, *aufzusetzen*, *aufsetzt*, *aufsetzte* sind die fünf häufigsten Formen) ist ersichtlich, dass jetzt nur diejenigen Formen analysiert werden, die als ein Wort (ununterbrochene Zeichenkette) vorkommen. Formen mit abgetrenntem Verbzusatz *auf* werden nicht berücksichtigt.

Ergebnisse anklicken

Durchführung der Kookkurrenzanalyse:

Kookkurrenzanalyse	anklicken

Dann eröffnet sich das Fenster, in dem die Parameter der Kookkurrenzanalyse eingestellt werden:

Einstellungen (Standardeinstellung)	
Kontext	5 Wörter rechts 5 Wörter links
Granularität:	grob
Zuverlässigkeit:	normal
Clusterzuordnung:	eindeutig
Lemmatisierung verwenden:	abwählen
Funktionswörter ignorieren	auswählen
LLR-Wert anzeigen:	auswählen
Nummerierung des Hauptkollokators:	auswählen
	Starten anklicken

Die Standardeinstellung (wie hier) ist empfehlenswert. Die Berechnung liefert „normale" Kookkurrenzpartner. Erläuterungen zu den weniger verständlichen Schaltern beschreiben Perkuhn/ Belica (2004) und Belica (1995) folgend:

„Autofocus: (Typische Stellung der Kollokatoren im Kontext ermitteln?)
Ohne Autofocus wird der gesamte eingestellte Kontext betrachtet, mit Autofocus werden alle möglichen Kontexte innerhalb des vorgegebenen Kontextes ausgewertet und es wird derjenige ausgewählt, der den höchsten Signifikanzwert aufweist

Granularität: (Wie intensiv sollen Mehrwortgruppen gesucht werden?)
Die Granularität gibt an, wieviele der nach Signifikanz sortierten Kookkurrenzpartner als möglicher Kandidat eines Kookkurrenzpartners n. Stufe in Frage kommen:
fein betrachtet die meisten (alle, die unter einem internen Schwellwert liegen) [zielt auf Wortverbindungen]
mittel betrachtet weniger (alle, die unter dem Schwellwert - 10 % liegen)
grob betrachtet am wenigsten (nochmals - 10 %) [zielt auf Schlagwörter]

Zuverlässigkeit: (Ziehen Sie Ausbeute oder Zuverlässigkeit vor?)
Inwieweit die Abweichung "beobachtet vs. normal" als relevant eingestuft werden soll, kann in drei Abstufungen vorgegeben werden:
hoch: nur starke Abweichungen sind relevant [findet wenige Kookkurrenzpartner, aber diese zuverlässig, ignoriert aber evtl. interessante Kandidaten, z.B. zufällig aufgrund Korpusauswahl und -komposition]
normal: mittlere Abweichungen sind relevant
analytisch: schwache Abweichungen sind relevant [findet viele Kookkurrenzpartner, aber diese evtl. unzuverlässig, kann u.U. auch schlechte Kandidaten mit erfassen]

Clusterzuordnung: (Was tun, wenn ein Beleg mehreren Kollokationsclustern zugeordnet werden kann?)
Bei "eindeutig" werden Ambiguitäten zugunsten des stärksten Kollokationsclusters aufgelöst.
Bei "mehrfach" werden Belege in alle relevanten Kollokationscluster eingefügt."

Ist die Funktion Lemmatisierung verwenden ausgewählt, erscheinen nur die Kookkurrenzpartner/ Kollokatoren (Abb. 44). Bei nicht eingeschalteter Lemmatisierung zeigt das Programm auch „syntagmatische Muster" (Abb. 45). Diese lassen sich durch Klicken auf ⊞ zu Konkordanzen umwandeln.

#	Total	Anzahl	LLR	Kookkurrenzen
1	5	5	824	**Krone**
2	2078	108	278	**Krönchen**
3	2247	169	247	**Sahne**
4	2813	566	161	**Brille**
5	2874	61	137	**Poker**

Abb. 44: Kookkurrenzen zur Abfrage: Lemma *aufsetzen*, lemmatisiert (Belica 1995, DeReKo)

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
1	1923	1923	1166	**Krone**	36% die Krone [...] aufgesetzt
2	2031	108	278	**Krönchen**	40% das|ein Krönchen [...] aufsetzen
3	2505	474	238	**Brille**	38% die|eine Brille [...] aufsetzen
4	2663	158	224	**Sahnehäubchen**	7% noch ein|das Sahnehäubchen [...] aufgesetzt werden
5	2729	66	201	**Meisterkrone**	63% die Meisterkrone [...] aufsetzen
6	2875	146	174	**Kopfhörer**	39% die Kopfhörer [...] aufgesetzt und ...
7	2951	76	139	**Narrenkappe**	42% die Narrenkappe [...] aufgesetzt
8	3000	49	127	**WM-Krone**	34% die WM-Krone [...] aufgesetzt hatte
9	3092	92	123	**i-Tüpfelchen**	32% noch das i-Tüpfelchen aufsetzen
10	3139	47	98	**Pokerface**	40% sein|ihr Pokerface [...] aufgesetzt und

Abb. 45: Kookkurrenzen zur Abfrage: Lemma *aufsetzen*, Wortformen (Belica 1995, DeReKo)

Die syntagmatischen Muster sind eigentlich sprachliche Chunks. In Anbetracht der Tatsache, dass sie einen hohen Signifikanzwert aufzeichnen, sollten sie auch im Fremd- und Zweitsprachenunterrich berücksichtigt und eingesetzt werden.

Bemerkungen:
Will man die Kookkurrenzpartner oder syntagmatischen Muster exportieren, muss beim Export → Ergebnisansicht der Posten Kookkurrenzen ausgewählt werden.

Die Kookkurrenzanalyse (in anderen Korpora auch Kollokationsanalyse genannt) ist eine Funktion, zu deren Ergebnissen ein Mensch, der nicht Jahre mit Recherchieren und Rechnen verbringen möchte, ohne ein elektronisches Korpus kaum gelangen könnte.

4.2.2 Korpora mit gesprochenem Deutsch

Repräsentative und große Korpora der gesprochenen Sprache sind viel schwieriger zu erstellen als die der geschriebenen Sprache. Dennoch existieren auch auf diesem Gebiet einige Quellen: **Bayerisches Archiv für Sprachsignale** (BAS); Österreichisches Aussprachewörterbuch/ **Österreichische Aussprachedatenbank** (adaba), **Korpus Gesprochene Sprache** beim DWDS, die **Datenbank für gesprochenes Deutsch** (DGD) oder das **GeWiss**-Korpus der gesprochenen Wissenschaftssprache. Darüber hinaus gibt es verschiedene Kleinkorpora mit unterschiedlichen Zielsetzungen (z.B. das Bochumer Korpus der gesprochenen Sprache im Ruhrgebiet, oder diverse Projekte am Hamburger Zentrum für Sprachkorpora). Keines von diesen Instrumenten ist so groß sowie stilistisch und regional ausgewogen, dass es für die gesprochene deutsche Sprache wirklich repräsentativ wäre. Das größte von ihnen ist die DGD.

Die meisten Korpora beinhalten standardisierte Transkripte der gesprochenen Sprache, zu denen die entsprechende Tonaufnahme synchronisiert abrufbar ist.

4.2.2.1 Datenbank für gesprochenes Deutsch (DGD)

http://dgd.ids-mannheim.de/
Die DGD ist momentan die größte Datenbank der gesprochenen deutschen Sprache (über 7,5 Mio. laufender Wörter), in der ähnlich wie in einem Korpus der geschriebenen Sprache recherchiert werden kann.
Für die Recherche benötigt man eine Registrierung. (Obwohl die DGD auch vom IDS-Mannheim verwaltet wird, gilt für sie nicht die Registrierung zu Cosmas II!)

Die Recherche ist relativ einfach, zuvor ist es jedoch ratsam, auf der übersichtlichen „Herzlich-Willkommen-Seite“ einen kurzen Blick auf die Beschreibungen der einzelnen Teilkorpora zu werfen.

Zugang zur DGD-Schnittstelle

Ins Google- oder Adressen-Suchfeld **http://dgd.ids-mannheim.de/**	eingeben
Registrierung	registrieren
Login	Anmelden
Recherche (auf der oberen Leiste) → Volltext	anklicken
Korpuswahl (am linken Rand)	aus-/abwählen

Eingabe der Abfrage:

Volltext-Suche

Beim Schreiben der Sucheingabe ins Suchfeld muss man immer den **Abstand zwischen den einzelnen Wörtern/ Zeichenketten und Sonderzeichen beachten!** Leerzeichen werden hier als ein Mittelpunkt (·) gekennzeichnet.
Die Möglichkeiten der Abfrage sind grundsätzlich diese:
Suche nach:

1) Einzelnen Wörtern bzw. Zeichenketten, die ein Bestandteil eines Wortes sind;
2) Wortkombinationen;
3) Kombinationen mit Abstand.

Abfragen: Wörter und Wortteile

Eingabe	Ergebnis
Arzt	Nur die Wortform *Arzt*
Arzt%	Alle Wörter, die mit *Arzt-* beginnen: *Arzt, Arztpraxis, Arztbesuche, Arztkoffer* sind die häufigsten in der DGD.
%arzt	Alle Wörter, die auf *-arzt* enden: *Notarzt, Hausarzt, Pseudoarzt, Frauenfacharzt, Dienstarzt* sind die häufigsten in der DGD.
_ _ _ _ _ig	Alle Wörter, die auf *-ig* enden und genau aus 6 Buchstaben bestehen: *fertig, völlig, witzig, häufig* sind die häufigsten in der DGD.

Tab. 10: Abfragen DGD (Wörter und Wortteile)

Erklärungen und Bemerkungen:

Symbol	Bedeutung
%	„Platzhalter“: öffnet die Grenze der Zeichenkette um eine beliebige Anzahl an Zeichen
_	„Platzhalter“: öffnet die Grenze der Zeichenkette um genau ein beliebiges Zeichen.

Abfragen: Wortkombinationen

Der Abstand (ein Leerzeichen) ist mit einem Mittelpunkt (·) markiert.

Eingabe	Ergebnis
kein·Spaß	Genaue Verbindung von Wörtern: *kein Spaß*
kein%·Spaß	Verbindung *kein Spaß*, *kein* in unterschiedlichen Formen: *kein, keinen, keinerlei*
groß&spaß	Alle Texte, in denen die Wörter *groß* und *Spaß* vorkommen.
lach%|läch%	Alle Texte, in denen eines oder beide Wörter, die auf *lach-* oder *läch-* beginnen, vorkommen: *lächelnd, lachen.*
lach%&spaß	Alle Texte, in denen *lachen, lachst, lacht* etc. und *Spaß* vorkommen: *spielen und lachen miteinander…und haben alle viel spaß*
$jung	*jung* lemmatisiert (alle flektierten Formen von *jung*): *jung, jüngeren*, aber auch *Junge, Jungs*

Tab. 11: Abfragen DGD (Wortkombinationen)

Erklärungen und Bemerkungen:

Symbol	Bedeutung
&	bedeutet „gleichzeitig“
|	bedeutet Alternative
$	„Lemmatisator“: sucht alle Formen des Grundwortes. (Kein Leerzeichen zw. $ und der Grundform)

Abfragen: Kombinationen mit Abstand

Suchanfrage	Ergebnis
NEAR(($geh,Schule),4,false)	Zwei Wörter im Abstand von max. 4 Positionen voneinander entfernt, Reihenfolge beliebig. (hier: … *in die **Schule gegangen**. Und **gehst** hier auch zur **Schule**?*)
NEAR(($geh,Schule),2,true)	Zwei Wörter im Abstand von max. 2 Positionen voneinander entfernt in dieser Reihenfolge. (hier: *Zwei **gehen** in die **Schule** …; Denn wir **gingen** mit der **Schule** …; Und dann **ging** ich zur **Schule**.*),

Tab. 12: Abfragen DGD (Kombinationen mit Abstand)

Erklärungen und Bemerkungen

Symbol	Bedeutung
NEAR	bedeutet Abstand (Keine Leerzeichen nach Kommas!)
Ziffer (hier: 4 und 2)	bedeutet Abstand (Keine Leerzeichen nach Kommas!)
false/true	Reihenfolge: false = beliebig; true = wie angegeben

Als Ergebnis bekommt man Konkordanzen der Transkripte, die KWICs sind färbig gekennzeichnet.

Weitere Funktionen:

Metadaten

Unter dem Menüpunkt Metadaten (auf der oberen Leiste unter Recherche → Metadaten) ist eine strukturensitive Metadatensuche möglich. Das heißt, dass die Audiofiles (und Transkripte) sich nach unterschiedlichen Kriterien einschränken und abrufen lassen: etwa nach Region, aus der die Sprecher/-innen kommen, Geschlecht, Dauer der Aufnahme, Anzahl der Sprecher/-innen etc.

Tokens

Unter Tokens (auf der oberen Leiste unter Recherche → Tokens) findet man die Abfragemöglichkeiten: **Wort**, **Normalisiert**, **Lemma**.

Wort heißt hier transkribiertes Wort, „lautgetreu" (z.B. [ˊvuʀʃt] = *wurscht*).

Normalisiert bedeutet der genormten Schreibweise angepasste Form (*wurst* oder *Wurst*).

Lemmatisiert ist eine Funktion, die alle Formen, die dem gewünschten Lemma zugeordnet wurden, abruft.
Die Abfrage nach *Wurst* (bitte Groß-/Kleinschreibung beachten!) ergibt:

Ergebnisse 1 bis 8 von 8 (0 ausgefiltert)				
	Ereignis		Treffer	
1	FOLK002	nein antwortet marmelad auf	wurst	hätt ich jetz mehr lust
2	FOLK021	nee wä malaria wär mir	wurscht	gewesen aber seine freundin schlägt der
3	FOLK024	mir können s ja auch mal sammlen is ja	wurscht	in welcher äh
4	FOLK024	in nem normale haushalt is is eigentlich ah	worscht	ob jetz gebortstag hosch oder net du mus
5	FOLK024	is ja irgendwie so die gschicht des is doch	wurscht	ob wir jetz enttäuscht wäre oder net entt
6	FOLK066	ah der eine der net warte konnt bis drei	würscht	roh esse also ich bitt dich
7	OS-009	dort mußten die Hochzeitsleute alle selber	Wurst	zu essen kaufen, ein Stückchen Wurst
8	OS-028	ds dann haben wir den Verwandten immer	Würste	und Fleisch hingetragen noch. Man hat's j

Abb. 46: Konkordanzen zur Abfrage: Lemma *Wurst* (DGD2)

Zu jedem Beispiel lassen sich die Tonaufnahmen und die Transkripte abrufen und herunterladen.

Das Korpus kann für unterschiedliche phonetische Übungen eingesetzt werden. Da zu den Transkripten Sequenzen von jeweils 15 Sek. (WMA-Format) abgerufen werden können,

könnte man die Transkripte und Tonsequenzen mischen und zuordnen lassen. Dies geht in einem E-Learning-Portal mit ein wenig Vorwissen ziemlich einfach.
Es eignet sich natürlich auch sehr gut für sämtliche wissenschaftliche Studien in der Phonetik und in der modernen Dialektologie.

4.2.2.2 DWDS - Gesprochene Sprache

http://retro.dwds.de/
Eine weitere bemerkenswerte Quelle für das gesprochene Deutsch ist das Korpus Gesprochene Sprache beim DWDS. Mit über 2,5 Mio. Tokens deckt das Korpus das ganze 20. Jahrhundert ab – von Aussagen Kaiser Wilhelms bis zu Sitzungsprotokollen des Deutschen Bundestags aus dem Jahr 1999. Es handelt sich zwar nur um standardisierte Transkripte ohne Tonspur (die Aussprache wurde dem üblichen Schriftbild angepasst – siehe Abb. 47), dennoch ist es eine ausgezeichnete Quelle für Recherchen zu gesprochenem Deutsch.

Die Abfragemöglichkeiten sind ident mit anderen DWDS-Korpora (siehe in Kap. 4.2.1.2)
Beispiel: In der Abb. 47 wurde die Wortform *net* abgerufen.

Suchfeldeingabe: @net

Corpus: gesprochene Sprache

Abfrage: @net #less_by_date[1900-01-01,2001-12-31] #cntxt 1 :spk

Trefferanzahl: 10 .

Seite: **1**

1	1938	STRE	… Scheißkerl ist, das ist ja ganz klar,	**net**	wahr….
2	1938	STRE	… sind wir doch ganz ehrlich,	**net**	wahr. …
3	1938	STRE	… nach München rüberfahren können,	**net**	wahr, da Kaffee trinken können….
4	1938	STRE	… bei uns wieder das Arbeiten,	**net**	wahr. Es ist gar keine Schande, wenn wir …
5	1938	STRE	…, wir sind doch beweglich,	**net**	wahr (Zwischengelächter), aber …
6	1989	JA	…. Ich kann (weiß i	**net**	) ich hab alle alle diese Sachen schriftlich, …
7	1989	BA	… und da hat man ja Kraft in der Gefahr,	**net**	wahr? Und da war ich mit, …
8	1989	CB	…, ich hab ja das	**net**	gewußt, einen Sauerteig gemacht, …
9	1994	AD	…, ich mach das	**net**	. Ich hab das selbst …

Seite: **1**

Abb. 47: Konkordanzen zur Abfrage: Wortform *net* (DWDS – gesprochene Sprache)

Da sich das Korpus momentan (2014) auf der alten Homepage des DWDS befindet, ist eine gesonderte Registrierung empfohlen (jedoch nicht notwendig).

4.2.3 Korpora mit historischen deutschen Texten

Studierende von DaF/DaZ sind meist Germanistikstudenten/-innen und haben oftmals eine germanistische Vorbildung. Im Studium, für die Forschung und auch als Demonstration der Vielfalt deutscher Sprache können auch historische und diachrone Korpora behilflich sein. In der Größe und Ausgewogenheit werden sie nie das Niveau von Korpora der gegenwärtigen Sprache erreichen können, aus naheliegenden Gründen wird es nie Korpora der gesprochenen Sprache aus der Zeit geben, in der Tonaufnahmen noch sehr rar waren oder gar nicht existierten. Dennoch ergänzen die Korpora mit historischen Texten das Bild über die deutsche Sprache um einen wesentlichen Aspekt.

Historische und diachrone Korpora werden oft gemeinsam mit synchronen Korpora verwaltet und bilden das „Nationale Korpus“. (z.B. Das Korpus Diakorp im Tschechischen Nationalkorpus).
Online gibt es mehrere Korpora, die nahezu alle überlieferten Stadien des Deutschen abdecken.

4.2.3.1 Deutsch Diachron Digital (DDD)

http://www.deutschdiachrondigital.de
Die Autoren schreiben über das Korpus des Althochdeutschen (DDD 2011):

> „Das Referenzkorpus Altdeutsch erfasst und annotiert sämtliche althochdeutschen und altniederdeutschen Texte (ca. 750 – 1050 u.Z.). Nach Abschluss der Arbeiten wird das Korpus etwa 650.000 Wörter umfassen.“
> Die Grundlage für das Korpus bilden als Referenztexte die handschriftengetreuesten gedruckten Texteditionen, soweit diese zugänglich sind. Alle Wortformen werden sowohl auf der Wortebene als auch auf der Buchstabenebene aufgenommen. Neben den Formen der Editionen werden die Schreibweise der Handschriften ebenso wie einheitlich standardisierte Wortformen in das Korpus aufgenommen. Graphische Besonderheiten wie z.B. Rubrizierung in den Manuskripten, Kursivierung, in den Editionen aufgelöste Diakritika oder Ligaturen werden im Korpus kommentiert.
> Alle Wortformen sind lemmatisiert. Die linguistische Annotation umfasst die Wortarten, morphologische Informationen und Angaben zu den Sätzen, des Weiteren werden Zeilenumbrüche, Absätze und andere Mittel zur Textgliederung ebenso wie Angaben zu Versgliederung und Reimpositionen, soweit vorhanden, erfasst.
> Alle in die Datenbank aufgenommenen Texte sind mit Header-Informationen versehen, die die sprach- und literaturwissenschaftlich relevanten Informationen zum jeweiligen Text wie z.B. Entstehungszeit, sprachlicher Raum und Überlieferungskontext enthalten.

Momentan läuft das Korpus unter dem Korpusmanager ANNIS (Search and Visualization in Multilayer Linguistic Corpora). Beschreibungen der Abfragen sind unter https://korpling.german.hu-berlin.de/annis3/ zu finden.

4.2.3.2 Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

http://mhdbdb.sbg.ac.at
Das Ziel des Projektes ist (MHDBDB 2012):

> „… vollständige mittelhochdeutsche literarische Texte elektronisch bearbeiten und speichern. [Dies] ergibt damit nicht nur die größte elektronische Sammlung deutscher Texte aus dem Mittelalter, sondern auch das bislang mächtigste und umfangreichste Abrufsystem für diese Sammlung.
> Fernziel ist es, die Textbasis weiter auszubauen und alle Texte voll-lemmatisiert und disambiguiert (d.h. nach Bedeutungskategorien aufgeschlüsselt) der Forschung über das Internet zur Verfügung zu stellen.“

Im Februar 2014 zählt die MHDBDB über 8 Mio. laufende Wörter in 418 Texten unterschiedlicher mittelalterlicher Quellen – von der Arthusdichtung bis zum Schwank und zu Verserzählungen (siehe MHDBDB → Textliste).

4.2.3.3 Bonner Frühneuhochdeutschkorpus (FNHD)

http://www.korpora.org/Fnhd/
Ein Korpus des Frühneuhochdeutschen entstand in Bonn: (FnhdC 2007):

> „Das Korpus besteht aus 40 Quellen, die nach Sprachlandschaften und Zeitschnitten (1350-1400, 1450-1500, 1550-1600 und 1650-1700) angeordnet sind. Es handelt sich um Auswahltexte mit einem Umfang von jeweils ca. 30 Normalseiten. Sämtliche Texte sind mit Wortklassenangaben und z.T. mit Formenbestimmungen annotiert."

4.2.3.4 Deutsches Referenzkorpus (DeReKo)

https://cosmas2.ids-mannheim.de/cosmas2-web/
Ein diachrones Korpus beinhaltet auch das DeReKo. Es „enthält Texte von der zweiten Hälfte des 17. Jahrhunderts bis 1972" (Cosmas II 2012) und befindet sich im Archiv HIST (Zugang über Cosmas II: Archiv → HIST-öffentlich).

Die letzte Version des Archivs HIST beinhaltet laut aktueller Abfrage (3.4.2014) Texte aus den Jahren 1650 bis 1979 und umfasst knapp 66 Mio. laufende Wörter in fast 5000 Texten (Abb. 48).

© Institut für Deutsche Sprache, Mannheim
COSMAS II-Server, C2API-Version 4.5.4 - 17. Feb. 2014

```
Datum          : Freitag, den 4. April 2014, 9:13:51
Korpus         : HIST-öffentlich - alle öffentlichen Korpora des Archivs HIST
Archiv-Release: Deutsches Referenzkorpus (DeReKo-2010-II)
```

Zusammensetzung des aktiven Korpus

Korpus-Ansicht, 10 Einträge, nach »bis« aufsteigend sortiert.

Texte	T(%)	Wörter	Jahrgänge	Korpus
795	16.195%	426.236	1816-1819	GRI Brüder Grimm: Sagen, Kinder- und …
29	0.591%	1.414.095	1772-1828	GOE Goethe-Korpus
680	13.852%	1.491.167	1833-1871	meg Korpus Marx-Engels-Gesamtausgabe
409	8.332%	2.444.587	1737-1877	KHZ Mannheimer Korpus Historischer Zeitungen
462	9.411%	825.950	1808-1882	mew Korpus Marx-Engels-Werke (ausg. Texte)
243	4.950%	1.649.049	1834-1905	KHM Mannheimer Korpus Hist. Zeitschriften
396	8.067%	14.351.045	1650-1927	HK4 Digitale Bibliothek: Dt. Lit. von Frauen
1.663	33.877%	30.298.776	1745-1927	HK3 Deutsche Lit. von Lessing bis Kafka
8	0.163%	168.277	1957-1962	mwa Anmerkungstexte zum Korpus Marx-Engels
224	4.563%	12.837.251	1713-1979	HK5 Philos. von Platon bis Nietzsche
4.909	**100.000%**	**65.906.433**	**1650-1979**	**10 Korpora**

Abb. 48: Zusammensetzung des Archivs HIST (DeReKo)

4.2.4 Parallelkorpora mit Deutsch

Die deutsche Sprache ist in mehreren Parallelkorpora vertreten, die unterschiedliche Größe, Eigenschaften und Recherchemöglichkeiten anbieten. Da für eine seriöse kontrastive Arbeit viele Belege aus unterschiedlichen Texten und zumindest die grundlegenden statistischen Werkzeuge benötigt werden, beschränkt sich die Vorstellung der Parallelkorpora mit Deutsch lediglich auf das OPUS-Corpus und das InterCorp. Beide bilden eine solide Basis nicht nur für verschiedene kontrastive Recherchen in zwei und mehreren Sprachen, sondern fungieren auch als einsprachige Korpora.

4.2.4.1 OPUS Korpus

http://opus.lingfil.uu.se/
Eine gute Quelle paralleler Texte bietet das „Open parallel corpus“ (kurz OPUS oder OPUS-Corpus), erstellt und sukzessive erweitert von Jörg Tiedemann (d.Z. an der Universität Uppsala).
Es beinhaltet Texte, die ausschließlich öffentlich zugänglich sind, aus dem Internet heruntergeladen und automatisch annotiert wurden (vgl. Tiedemann 2012: 2214). Die meisten Texte, in denen man mit Korpustools browsen kann, sind Fachtexte (Details in Tiedemann (2009)). Das Korpus beinhaltet für die Erkundung der gesprochenen Sprache auch eine wertvolle Sammlung von Film-Untertiteln (dazu Tiedemann 2007).
Das Korpus besteht aus mehreren Sub-Korpora, in denen recherchiert werden kann (siehe **Sub-corpora (downloads & infos)** auf der Hauptseite

Unter **Search & download resources** auf der Hauptseite kann man sich ein Bild davon machen, in welchen Subkorpora die gewünschten Parallelen zu finden und wie groß diese sind.

Die Abfragemöglichkeiten sind eingeschränkter als beim InterCorp, dennoch im Wesentlichen ident (dies betrifft die Syntax der Abfrage, das Filtern, Frequenzen etc.). Aus diesem Grund wird auf eine detaillierte Beschreibung der einzelnen Tools im OPUS-Corpus verzichtet.

Es eignet sich auch als eine hervorragende Ergänzung zum InterCorp, weil hier Recherchen in einigen Sprachenkombinationen möglich sind, die im InterCorp noch nicht existieren: z.B. Deutsch – Türkisch, Deutsch – Chinesisch, Deutsch – Japanisch (Abb. 49).
Beispiel: Gesucht wird im Deutschen – Japanischen nach Wörtern, die im Deutschen auf *Haus-* beginnen:

Abfrage:
Search & Browse → OPUS multilingual search interface → OpenSubtitles → de

ja (= Japanisch)	auswählen
vertical	auswählen

Eingabe: **[word="Haus.*"]**
select anklicken

Die Ergebnisse der Abfrage sind parallele Konkordanzzeilen – Sätze bzw. Repliken der Filmfiguren:

Query string: '[word="Haus.*"] :
20 hits found

	de	ja
413066	Ich habe ein **Haus** in Brentwood .	私 の 家 は ブレントウッド です
413079	' in mein **Haus** eingebrochen und haben meine **Haushälterin** getötet . '	家 に 押し入っ て 来 て 家政 婦 を 殺し た の
414258	Ja , zu **Hause** , Jessica .	ああ 家 に ね
418314	Ich werd dich nach **Hause** bringen .	わかっ たよ ママ
1095165	Die **Haustür** stand letzte Nacht wieder offen .	昨日 また 玄関 開けっ放し だっ た ぞ
1752181	Zu **Hause** .	また ね
1752701	Ich führe zurzeit zwei **Haushalte** .	２ 世帯 を 世話 し てる ので
1756657	Komm einfach abends nach **Hause** .	夜 まで に は 帰っ てき て ―

Abb. 49: Parallele Konkordanzen zur Abfrage *Haus-* (OPUS-Corpus, Open Subtitles: Deutsch – Japanisch)

Die komplette Seite, Beschreibungen und die Bedienung des OPUS-Corpus sind auf Englisch.

4.2.4.2 InterCorp

http://ucnk.ff.cuni.cz/intercorp/
InterCorp besteht aus parallelen Korpora mehrerer Sprachen. Dieses multilinguale Korpus bildet einen Teil des Projektes „Das Tschechische Nationalkorpus und Korpora anderer Sprachen“. Die Basis des InterCorp wurde in den Jahren 2005 bis 2011 aufgebaut und beinhaltete damals 22 Sprachen. Mittlerweile ist die Recherche in 33 Sprachen möglich und weitere, auch außereuropäische Sprachen sollen ergänzt werden. Weitere Informationen über das Korpus in deutscher Sprache sind auf der deutschen InterCorp-Homepage zu finden (Káňa/ Vavřín: 2011). Momentan liegt das Manual für das InterCorp nur auf Tschechisch, in Zukunft auch auf Englisch vor. Eine deutsche Fassung ist nicht geplant, weshalb hier die wichtigsten Arbeitsanweisungen dargestellt werden:

Das Ziel des Projektes ist die Erstellung von synchronen Parallelkorpora der tschechischen und jeweils einer anderen Sprache. So sollen akademische und nicht kommerzielle Parallelkorpora mit Tschechisch und den meisten Fremdsprachen, die an den geisteswissenschaftlichen Fakultäten in Tschechien studiert werden, entstehen.

Das Korpus InterCorp nimmt in mehreren Hinsichten eine Sonderrolle unter den Korpora des Instituts fürs Tschechische Nationalkorpus ein. Die wesentlichen Unterschiede sind folgende:

- Jede Parallele fungiert als ein vollwertiges einsprachiges Korpus mit allen zugänglichen Instrumenten (Filtern, Sortieren, Verteilung…)
- Im Unterschied zu Referenzkorpora zeichnet sich das InterCorp durch einen stetigen Zuwachs im Umfang der Paralleltexte und auch in der Anzahl der Sprachen aus. Es kann aber nicht automatisch als Referenzkorpus verwendet werden, da die Texte weder stilistisch noch regional noch originalsprachlich (nach Richtung der Übersetzung) ausgewogen sind.

Die Parallelkorpora dienen als Datenquellen für theoretische Studien, studentische Arbeiten, für die Lexikographie und vor allem auch als Unterstützung für den Fremdsprachenunterricht.

Weiters stehen sie auch Übersetzer/-innen und der breiten Öffentlichkeit zur Verfügung. In der Anbahnungsphase des Projektes wurden die Parallelkorpora an der jeweils zuständigen Arbeitsstelle erstellt und ausprobiert. In weiteren Phasen wurden sie auf einem zentralen Server zusammengeführt und veröffentlicht.

Hinweise zur Nutzung von InterCorp
Nach der Registrierung auf der Hauptseite (inkl. einer unbürokratischen Zustimmung zu den Nutzungsbedingungen) ist der Zugang über das Web-Interface frei. Bereits registrierte Nutzer des ÚČNK/ICNC haben automatisch Zugriff auch auf das InterCorp.

Registrierung:
www.korpus.cz → **Agreements and registration/ Registrace nového uživatele** → **Statements of a User of the ICNC Corpora** (Erklärung des Benutzers der ÚČNK-Korpora) (nur Englisch)

4.2.4.2.1 Texte und Sprachen des InterCorp

Im InterCorp sind mehrheitlich manuell alignierte, d.h. in den Parallelen sich entsprechende Texte gespeichert. Laufend werden auch Texte aus *Project Syndicate* und *Presseurop* eingespeist und zugänglich gemacht. Es ist jedoch darauf hinzuweisen, dass diese weiteren Texte lediglich automatisch aligniert werden. Die Konkordanzen (Trefferzeilen) können daher in ihren Entsprechungen Fehler aufweisen.
Jeder Text im Korpus hat auch (s)ein tschechisches Pendant, mit dem er primär aligniert ist. Im Zentrum steht also jeweils ein Text auf Tschechisch, zu dem es eine oder mehrere fremdsprachige Variante(n) gibt. Der Gesamtumfang des InterCorp ändert sich mit jeder Version und ist auf der Homepage festzustellen. Die Version 6 (veröffentlicht am 8. 4. 2013) hatte 68,509.857 Positionen bzw. 56,835.000 Wörter in der deutschen Sprache (Dovalil/ Káňa/ Pelouškova et al. 2013).

Deutsch kann man kontrastiv mit allen im InterCorp vertretenen Sprachen untersuchen. Das bedeutet, dass es jeweils mindestens einen Text gibt, der auf Deutsch und einer weiteren Sprache im InterCorp enthalten ist. Allerdings muss man damit rechnen, dass zu einigen Sprachen nur ganz wenige Paralleltexte vorhanden sind. Etwa die Sprachkombinationen Deutsch – Arabisch oder Deutsch – Hindi ermöglichen derzeit eine Recherche in nur einem oder zwei Texten.

Die Tabellen auf der folgenden Seite charakterisieren das InterCorp, Version 6. Sie präsentieren eine alphabetische Übersicht über die Sprachen in einzelnen Parallelen (Tab. 13) und eine Übersicht über die aktuelle Größe der jeweiligen Sprachparallele (Tab. 14). Die Angaben zur Größe werden in 1.000 Wörtern angeführt.

Parallelen im InterCorp - Übersicht

Kürzel	Link	Sprache	Wörter 10^3
ar	intercorp_ar	**Arabisch**	29
bg	intercorp_bg	**Bulgarisch**	26 879
da	intercorp_da	**Dänisch**	35 785
de	intercorp_de	**Deutsch**	56 835
en	intercorp_en	**Englisch**	54 753
et	intercorp_et	**Estnisch**	26 862
fi	intercorp_fi	**Finnisch**	29 040
fr	intercorp_fr	**Französisch**	53 936
el	intercorp_el	**Griechisch**	40 683
hi	intercorp_hi	**Hindi**	155
it	intercorp_it	**Italienisch**	46 560
ca	intercorp_ca	**Katalanisch**	1 758
hr	intercorp_hr	**Kroatisch**	12 625
lv	intercorp_lv	**Lettisch**	31 770
lt	intercorp_lt	**Litauisch**	29 811
mt	intercorp_mt	**Maltesisch**	14 133
mk	intercorp_mk	**Mazedonisch**	2 664
nl	intercorp_nl	**Niederländisch**	51 817
no	intercorp_no	**Norwegisch**	2 301
pl	intercorp_pl	**Polnisch**	47 640
pt	intercorp_pt	**Portugiesisch**	49 502
ro	intercorp_ro	**Rumänisch**	21 995
ru	intercorp_ru	**Russisch**	7 588
sr	intercorp_sr	**Serbisch (lat.)**	6 972
sy	intercorp_sy	**Serbisch (kyr.)**	2 443
sv	intercorp_sv	**Schwedisch**	41 694
sk	intercorp_sk	**Slowakisch**	40 108
sl	intercorp_sl	**Slowenisch**	33 741
es	intercorp_es	**Spanisch**	62 865
cs	intercorp_cs	**Tschechisch**	99 547
uk	intercorp_uk	**Ukrainisch**	1 493
hu	intercorp_hu	**Ungarisch**	33 985
be	intercorp_be	**Weißrussisch**	1 308

Tab. 13: Sprachen im InterCorp alphabetisch

Kürzel	Link	Sprache	Wörter 10^3
cs	intercorp_cs	Tschechisch	99 547
es	intercorp_es	Spanisch	62 865
de	intercorp_de	Deutsch	56 835
en	intercorp_en	Englisch	54 753
fr	intercorp_fr	Französisch	53 936
nl	intercorp_nl	Niederländisch	51 817
pt	intercorp_pt	Portugiesisch	49 502
pl	intercorp_pl	Polnisch	47 640
it	intercorp_it	Italienisch	46 560
sv	intercorp_sv	Schwedisch	41 694
el	intercorp_el	Griechisch	40 683
sk	intercorp_sk	Slowakisch	40 108
da	intercorp_da	Dänisch	35 785
hu	intercorp_hu	Ungarisch	33 985
sl	intercorp_sl	Slowenisch	33 741
lv	intercorp_lv	Lettisch	31 770
lt	intercorp_lt	Litauisch	29 811
fi	intercorp_fi	Finnisch	29 040
bg	intercorp_bg	Bulgarisch	26 879
et	intercorp_et	Estnisch	26 862
ro	intercorp_ro	Rumänisch	21 995
mt	intercorp_mt	Maltesisch	14 133
hr	intercorp_hr	Kroatisch	12 625
ru	intercorp_ru	Russisch	7 588
sr	intercort_sr	Serbisch (lat.)	6 972
mk	intercorp_mk	Mazedonisch	2 664
sy	intercorp_sy	Serbisch (kyr.)	2 443
no	intercorp_no	Norwegisch	2 301
ca	intercorp_ca	Katalanisch	1 758
uk	intercorp_uk	Ukrainisch	1 493
be	intercorp_be	Weißrussisch	1 308
hi	intercorp_hi	Hindi	155
ar	intercorp_ar	Arabisch	29

Tab. 14: Sprachen im InterCorp nach Größe

Morphosyntaktische Annotation im InterCorp

Im InterCorp wurden bisher Texte der folgenden Sprachen morphosyntaktisch annotiert:

Sprache	Taggi ng	Lemmatisi erung	Sprache des Manuals	Tagger
Bulgarisch	✔		**Englisch**	**TreeTagger**
Deutsch	✔	✔	**Deutsch**	**TreeTagger**
Englisch	✔	✔	**Englisch + Ergänzungen**	**TreeTagger**
Estnisch	✔	✔	**Englisch, Estnisch**	**TreeTagger**
Französisch	✔	✔		**TreeTagger**
Italienisch	✔	✔		**TreeTagger**
Litauisch	✔	✔		**von Vidas Daudaravičius**
Niederländisch	✔			**TreeTagger**
Norwegisch	✔	✔		**Oslo Bergen Tagger**
Polnisch	✔	✔	**Englisch**	**Morfeusz, TaKIPI**
Portugiesisch	✔	✔	**Spanisch**	**TreeTagger**
Tschechisch	✔	✔	**Englisch**	**Morče**
Russisch	✔	✔	**Englisch**	**TreeTagger**
Slowakisch	✔	✔	**Slowakisch**	**Radovan Garabík, Morče**
Slowakisch	✔	✔	**Slowakisch**	**Radovan Garabík, Morče**
Slowenisch	✔	✔	**Englisch**	**totale**
Spanisch	✔	✔	**Spanisch**	**TreeTagger**
Ungarisch	✔		**Englisch**	**HunPos**
Arabisch Dänisch Finnisch Griechisch Hindi Katalanisch Kroatisch Lettisch Maltesisch Mazedonisch Rumänisch Serbisch (kyr.) Serbisch (lat.) Schwedisch Ukrainisch Weißrussisch	**nicht getagget**			

Tab. 15: Übersicht der morphosyntaktischen Annotation im InterCorp (2014)

4.2.4.2.2 ARBEIT MIT DEM INTERCORP

Die Internetseite des Tschechischen Nationalkorpus, auf der sich auch das InterCorp befindet (www.korpus.cz), ist nur in englischer und tschechischer Sprache verfasst. Grundkenntnisse des Englischen sind demnach für die Arbeit mit diesem Korpus erforderlich.

Falls die Seite auf Tschechisch erscheint, kann man auf der oberen Leiste rechts auf English umschalten. Dann erscheinen auf der Leiste diese Posten:

KonText Park SyD Morfio KWords Login | New user registration česky

Eine Registrierung über New user registration ist erforderlich, jedoch schnell und unbürokratisch.

Nach der Registrierung auf KonText klicken, dann erscheint dieses Bild:

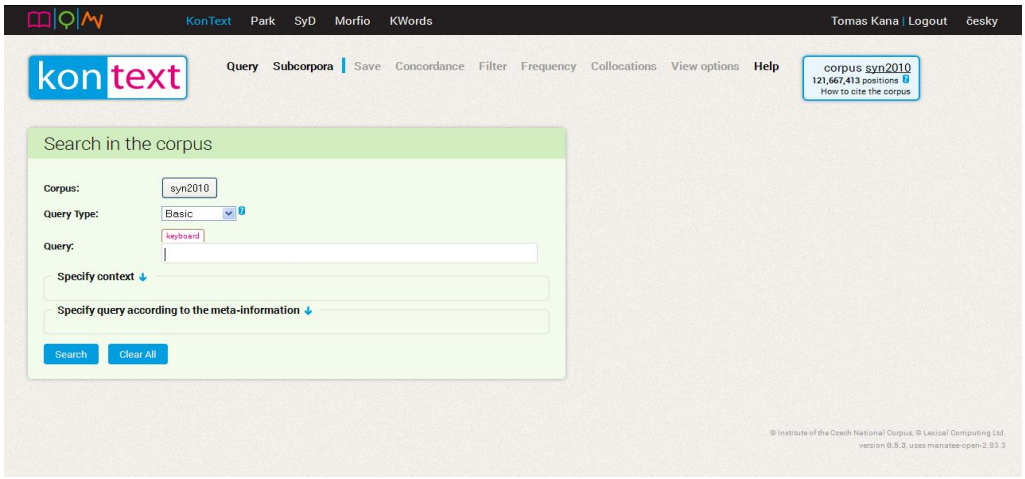

Abb. 50: KonText Startseite (Default: Korpus syn 2010)

Im grünen Viereck **Search in the Corpus** stellt man das gewünschte Korpus ein:
Dazu den grauen Button neben **Corpus:** anklicken (vorangestellt ist das Korpus **syn2010** - synchrones Korpus des Tschechischen, für DaF/DaZ irrelevant), dann erscheint folgendes Angebot:

- ▶ Synchronic written corpora
- ▶ Synchronic spoken corpora
- ▶ Diachronic corpora
- ▶ Foreign language corpora
- ▶ Foreign language web corpora
- ▶ Parallel corpus InterCorp

Für DaF/DaZ sind primär Korpora mit Deutsch relevant. Die findet man unter:

▶ Foreign language web corpora

Beim Anklicken des Texts Foreign language web corpora öffnet sich das Angebot an Web-Korpora. Hier ist u.a. auch das Korpus **deWaC** zu finden. Dieses beinhaltet Texte ausschließlich aus dem Internet, hat über 1,5 Milliarden Tokens/ Positionen (Stand: Jänner 2014) und dieselben Eigenschaften wie die deutsche Parallele im InterCorp.

▶ Parallel corpus InterCorp

Nach dem Klick auf Parallel corpus InterCorp öffnet sich das Angebot an sprachlichen Parallelen. Derzeit sind es Sprachen, die in den Tab. 13 und Tab. 14 angeführt sind.
Die Texte der jeweiligen Sprache haben immer ihr Pendant im Tschechischen und viele von ihnen auch in einer anderen Sprache. Es sind bereits viele Texte auf Tschechisch, Deutsch und noch in einer anderen Sprache im Korpus vorhanden. Man kann mit relativ großen Korpora (einige Mio. Wörter) mit Deutsch und noch einer anderen großen Parallelsprache arbeiten. Zu beachten ist jedoch Folgendes: je „kleiner" oder von Mitteleuropa aus gesehen entfernter die Sprache ist, desto weniger Texte sind im Parallelkorpus zu finden.
Aus dem Angebot der Parallelen wählt man jetzt Deutsch durch einen Klick auf intercorp_de aus.

Im Weiteren wird mit der deutschen Parallele (intercorp_de) gearbeitet. Die Eigenschaften (inkl. Tagging) sind für die Korpora **intercorp_de** und **deWac** ident. Sie decken sich sogar mit vielen Eigenschaften anderer deutschsprachigen Korpora (v.a. **DeReKo** und **DWDS**). Die Korpora **intercorp_de** und **deWac** unterscheiden sich vor allem in der Größe, Struktur der Texte und natürlich in der Möglichkeit, dieselben Texte in einer anderen Sprache abzurufen – dies ist im deWaC nicht möglich.

Beschreibung des InterCorp unter KonText

Das Korpus InterCorp und andere Korpora des Instituts (für das Tschechische Nationalkorpus) laufen unter mehreren Schnittstellen. Die neueste heißt KonText und ist im Prinzip der Schnittstelle (No)Sketch Engine (frühere Version des InterCorp) sehr ähnlich.
Die einzelnen Funktionen und Einstellungen sind einfach und intuitiv zu bedienen. Hier werden sie in drei Kapiteln (in der logischen Abfolge zu einer Abfrage) beschrieben:

A. Aktuelles Korpus
B. Abfragefenster
C. Einstellungen und Funktionen

Diese Kapitel widerspiegeln das Bild der Arbeitsfläche (Abb. 51). In ihnen werden die Bedeutungen einzelner Buttons und Funktionen erklärt und um Beispiele ergänzt.

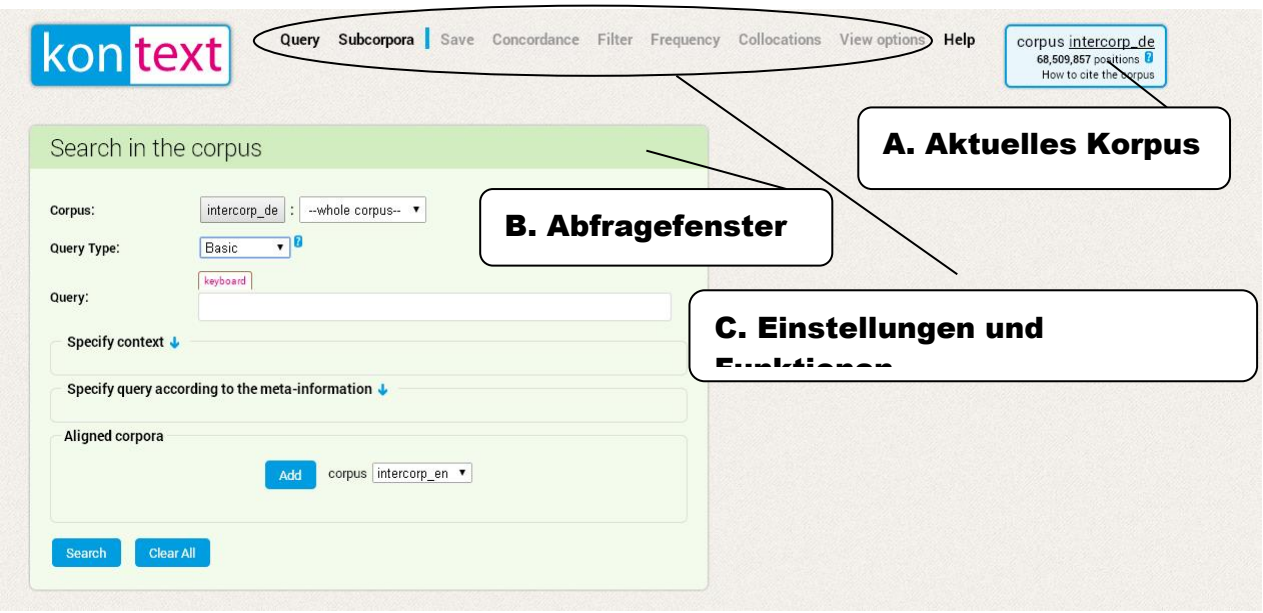

Abb. 51: Kontext-Startseite (InterCorp_de)

A. Aktuelles Korpus

corpus:	
intercorp_de:	Name des Korpus (hier intercorp_de = deutsche Parallele). Durch Anklicken dieses Buttons erscheinen ausführliche Informationen über diese Parallele (die gleichzeitig auch als ein vollwertiges Korpus funktioniert).

In der Version 6 (2013) sind es für Deutsch folgende Angaben:

Attributes		
word	885.504	Anzahl unterschiedlicher Wörter/ Positionen
lemma	706.237	Anzahl unterschiedlicher Lemmata einzelner Positionen
tag	54	Anzahl der Tags (s. Tagset)
lc	823.825	Anzahl der Wörter (ohne Satzdiakritika)
lemma_lc	677.282	Anzahl der Lemmata (Satzdiakritika ausgeschlossen)

Tab. 16: Technische Angaben: Attribute (InterCorp_de, Version 6)

Structures		
<doc>	231	Anzahl der Dokumente im Korpus
** **	89. 734	Anzahl struktureller Attribute (Informationen über Texte)
<p>	1,748.827	Anzahl der Absätze
<s>	3,714.390	Anzahl der Sätze

Tab. 17: Technische Angaben: Strukturen (InterCorp_de, Version 6)

68,509,857 positions **Gesamtgröße des Korpus in Positionen**
Durch Anklicken des Fragezeichens erscheint eine kurze Definition des Begriffes Position.

How to cite the corpus **Hinweise zu den Zitierregeln**
Durch das Anklicken bekommt man bibliographische Angaben, wie jegliche Belege und Daten aus dem Korpus zu zitieren sind. Bitte bedenken Sie, dass das konsequente und richtige Zitieren die Qualität jeder Arbeit erhöht und in der wissenschaftlichen Praxis zum Standard gehört. Die Autor/-innen sind für Hinweise über die Zitate dankbar.

B. Abfragefenster/ Search in the corpus

Corpus:	Name des aktuellen Korpus Falls ein Subkorpus erstellt wurde, erscheint nach dem Korpusnamen ein Fensterchen, in dem zur Auswahl steht, ob im gesamten Korpus (– whole corpus –) oder nur in einem Teil davon recherchiert wird. Vorangestellt (Default) ist – whole corpus – Zur Erstellung der Subkorpora: auf der oberen Leiste Subcorpora anklicken (Beschreibung hier im Abschnitt C)

Query Type:	Abfragemodus: im Auswahlfenster kann man verschiedene Abfragemöglichkeiten (Attribute der Abfrage) einstellen: **Basic** **Lemma** **Phrase** **Word Form** **Charakter** **CQL**

Query Type:	
Basic:	bedeutet „Einstiegsmodus". Gibt man eine Grundform („Wörterbuchform") ein (z.B. *lustig*), sucht das Korpus nach alle Formen des Wortes (*lustig, Lustig, lustigen, lustiger, lustiges, lustigste* etc.). Wenn eine flektierte Form (z.B. *Täler*, *lustiger*, *dich*, *möchte*), sucht das Korpus nur diese. Im Basic-Modus wird die Groß-/Kleinschreibung nicht beachtet.

Beispiel

Eingabe	**Ergebnis (Beschreibung)**
lustig	*lustig, Lustig, lustigen, lustiger, lustiges, lustigste …* (alle Formen)
Täler	*Täler, täler* (nur diese Formen)

Konkordanzen (Zufallsauswahl):

Und genau das mache ich .	**Lustig**	, was ? «.
Die Galaxis ist ' ne	**lustige**	Sache . Du musst dir bloß diesen Fisch
Es ist das	**Lustigste**	, was man sich überhaupt denken kann

Abb. 52: Konkordanzen zur Abfrage: (Basic) *lustig* (InterCorp_de)

denkt Klemmer an die Berge und	**Täler**	Österreichs ,
und der Dampfer pflügte sich (…) durch die Wellenberge und -	**täler**	.

Abb. 53: Konkordanzen zur Abfrage: (Basic) *Täler* (InterCorp_de)

Query Type:	
Lemma:	Abfrage nach einzelnen Formen einer Grundform. Ins Abfragefenster muss nur die Grundform[18] eingegeben werden (*Buch*, *lustig*, *schreiben*), anderenfalls zeigt das Korpus „Empty result“

Beispiel:

Eingabe	Ergebnis (Beschreibung)
schreiben	*schreiben*, *schreibe*, *schrieb*, *geschrieben* … (alle Flexionsformen)

Konkordanzen (Beispiele, Zufallsauswahl):

der Befehl , von einem General eigenhändig	**geschrieben**	: » Alle sind auf der Stelle zu
paar Wochen angefangen , eine Geschichte zu	**schreiben**	, etwas , das ganz ausgedacht ist , es
Ich erkundigte mich , was er	**schreibe**	. Er beschreibe sein Leben .
schwierig , aber nicht schlecht	**geschrieben**	. Der Verfasser dieser Beiträge aus
die astrologische Rubrik ihrer Zeitschrift zu	**schreiben**	, war ich
Versuche gelang . Dann	**schrieb**	er ihr einen Brief sowohl ins Büro als
sie auch mich . Peter	**schreibt**	. Peter sagt . Alles Lüge , Lüge , Lüge
Mannes , von dem ich noch nie gehört hatte ,	**geschrieben**	von einem Mann , von dem ich
Gleich in medias res gehen .	**Schreib**	! … Ich , Valerie Steinfeld , geborene

Abb. 54: Konkordanzen zur Abfrage: Lemma *schreiben* (InterCorp_de)

Query Type:	
Phrase:	Abfrage nach zwei und mehr Wörtern/ Zeichenketten, die durch ein Leerzeichen (hier durch einen Punkt „·“ gekennzeichnet) getrennt sind.

Eingabe	Ergebnis (Beschreibung)
rosarote·Brille	*rosarote Brille* (nur diese Formen)

Konkordanzen (Beispiele, Zufallsauswahl):

die Lage vor dem Eingreifen durch eine	**rosarote Brille**	zu sehen . Wie bereits betont
das Parlament bei einer derartig durch die	**rosarote Brille**	gesehenen Herangehensweise
Reis , Honig . “ Aber ich versuche , eine	**rosarote Brille**	zu tragen . ” Irgendwann werde
aber durch eine wesentlich weniger	**rosarote Brille**	. Nicht nur , dass der Begriff Kultur in

Abb. 55: Konkordanzen zur Abfrage Phrase: *rosarote Brille* (InterCorp_de)

Eingabe	Ergebnis (Beschreibung)
rosa.*·Brille	*rosa Brille*, *rosarote Brille*, *rosaroten Brille* … (alle Wörter auf *rosa-* gefolgt von *Brille*)

Konkordanzen (Beispiele, Zufallsauswahl):

bei einer derartig durch die	**rosarote Brille**	gesehenen Herangehensweise mitmacht
“ Aber ich versuche , eine	**rosarote Brille**	zu tragen . ”
„ Dann sage ich : nehmt die	**rosa Brille**	ab “ und spielt dabei auf
Wenn man dieses Bild mit einer	**rosaroten Brille**	betrachtet , so könnte man

Abb. 56: Konkordanzen zur Abfrage: Phrase *rosa- Brille* (InterCorp_de)

[18] Diese findet man als Wörterbucheintrag: bei Substantiven, Adjektiven etc. ist es die Form im Nominativ Singular (bei Adjektiven und Adverbien Positiv), bei Verben ist es die Form im Infinitiv Präsens.

Query Type:	
Word Form:	Abfrage nach einer bestimmten Form oder nach Teil eines Wortes. Die Eingabe muss eine ununterbrochene Kette von Zeichen sein (ohne Leerzeichen). Die Abfrage nach einem Teil erfolgt über Platzhalterzeichen (Tab. 18 auf der folgenden Seite).

Eingabe	Ergebnis (Beschreibung)
rosa	*rosa* (nur diese Form; Groß-/Kleinschreibung ausgeschaltet)

Konkordanzen (Beispiele, Zufallsauswahl):

und nährte gigantische	**rosa**	und blaue Hortensiensträucher ,
Herr Karl ihr ein	**rosa**	Netz über die Wickler legte ,
dazu tragen und ein	**rosa**	Hemd und eine gestreifte Krawatte
Rosa , schreibend an	**Rosa**	und Rosa begehrend ,
ein weißes und ein	**rosa**	Zettelchen heraus ,

Abb. 57: Konkordanzen zur Abfrage *rosa* (Word Form)

Eingabe	Ergebnis (Beschreibung)
.rosa.*	Ein beliebiger Buchstabe links, beliebige Anzahl beliebiger Buchstaben rechts: *Prosa*, *Brosamen*, *Arosa*, *prosaisch*, *prosaischen*, *Prosawerk* sind die häufigsten im InterCorp_de.

Konkordanzen (Beispiele, Zufallsauswahl):

stelle ich Ihnen die folgende	**prosaische**	Frage : Wie viel hat es Spanien
Auch in der	**Prosa**	gibt es kaum noch einen Autor ,
So lautete das Eingangsgedicht des	**Prosabandes**	" Frühling um 1900 « , das 1936
traurige Unzulänglichkeit aller	**Prosadichtung**	gegenüber der Aufgabe , die sie
ist später dann in dem erwähnten	**Prosatext**	festgehalten worden .
sein Schicksal , immer nur die	**Brosamen**	einsammeln zu können

Abb. 58: Konkordanzen zur Abfrage: Wortform -*rosa*- (InterCorp_de)

Eingabe	Ergebnis (Beschreibung)
.*rosa.*	Eine beliebige Anzahl beliebiger Buchstaben sowohl rechts als auch links: *Prosa, Brosamen…; blaßrosa, rosafarben(en), zartrosa* sind die häufigsten im InterCorp_de.

Konkordanzen (Beispiele, Zufallsauswahl):

der Zeit das Geld ersetzte . Mit den	**rosa**	Zettelchen wurde alles im Kramladen
Der Fahrer , der Bischof	**Aringarosa**	am Flughafen Leonardo da Vinci in Rom
In der Tat hätte die Bevölkerung	**Ombrosas**	, wäre die Sache ruchbar geworden , die
eingeschlossen , sind	**sakrosankt**	- tabu .
Sie hat den	**prosaischen**	Blick , wasserblau natürlich
Nasenflügel , legte einen	**zartrosa**	Lippenstift auf und musterte sich im
Edelsteins sprühten in einem	**rosafarbenem**	Feuer .
ihre langen Fingernägel waren	**knallrosa**	lackiert ,

Abb. 59: Konkordanzen zur Abfrage: Worform -*rosa*- (InterCorp_de)

Platzhalterzeichen und andere reguläre Ausdrücke im InterCorp:

	Symbol	Erklärung
.	Punkt	ein beliebiges Zeichen
*****	Asterisk	beliebige Anzahl an Zeichen
+	Pluszeichen	einmalige oder mehrmalige Wiederholung des vorhergehenden Zeichens: z.B. **Buf+et+** ergibt *Bufet*, *Buffet*, *Bufett*, *Buffett*)
[]	Eckige Klammern	umgeben Alternativen: z.B. **B[uü]ffet** ergibt *Buffet* und *Büffet*[19].
^	Zirkumflex	negiert die Alternative in der Klammer: z.B. **gr[^ü]ße** ergibt *größe*, *große*, jedoch nicht *grüße*.
()	Runde Klammern	geben Prioritäten an – wie beim Rechnen: z.B. **Schluß|ss** ergibt *Schluß* oder *SS*; während **Schlu(ß|ss)** ergibt *Schluß* oder *Schluss*.
{}	Geschwungene Klammern	geben Intervall der Wiederholungen an: z.B. **[tag="ADJA "]{3,5}** ergibt Adjektivhäufungen von 3 bis 5 Adjektiven hintereinander: *neuen globalen politischen* oder *ausgezeichneten neuen britischen parlamentarischen* sowie *europaweiten öffentlichen zellularen digitalen terrestrischen*
?	Fragezeichen	wiederholt das vorhergehende Zeichen oder die vorhergehende Zeichenkette einmal oder null-mal: z.B. **(ur)?urgroßeltern** ergibt *Ururgroßeltern* und *Urgroßeltern*.
!	Rufzeichen	negiert das darauffolgende (relevant für CQL-Abfrage): z.B. **[word=".*lein"&!word="allein"]** ergibt *Fräulein*, *Häuflein*, *Büchlein*, *klein* …
****	Backslash	muss eingesetzt werden, wenn nach einem der hier angeführten Sonderzeichen gesucht wird: z.B. **\+** sucht nach Pluszeichen im Text und ergibt *Schale mit der Aufschrift MILCH + ALKOHOL* oder *Canal +*

Tab. 18: Reguläre Ausdrücke im InterCorp

Shortcuts zu diesen Symbolen befinden sich im Kap. 8.3.

Query Type:	
Charakter:	Abfrage nach einer beliebigen Zeichenkette innerhalb eines Wortes. Dieser Modus eignet sich gut für die Erschließung von Wortfamilien.

Eingabe	**Ergebnis (Beschreibung)**
lust	ergibt alle Wörter, die die Buchstabenkombination (hier *lust*) beinhalten: *Verlust, Verlusten, Verlustrechnung , lustig, belustigt, lustlos, Wollust, Cluster, illustriert, Ballustrade* sind die 10 häufigsten im InterCorp_de.

[19] Nach Daten des Korpuses DeWaC ist übrigens *Buffet* (80%) etwa viermal üblicher als *Büffet* (20%).

Konkordanzen (Beispiele, Zufallsauswahl):

Selbstverwirklichung und dem	**Bedeutungsverlust**	kollektiver Interessensvertretung
Programm Perdita Durango Ein	**lebenslustiger**	Santería-Priester sucht
die Mini-Bar (hier drohte schließlich	**Umsatzverlust**	;-))repariert .
Sicherheitsgründen vor	**Verlust**	gelassen und konnten sich
Einen auch mit Bildern	**illustrierten**	Gang durch die Geschichte der
Bildung von räumlichen und sektoralen "	**Clustern**	" erfolgversprechend .
gesellten sich zu unserer	**lustigen**	Runde und demonstrierten ihr "
Auslagerung kann auch einen	**Verlust**	an Know-how und an Flexibilität

Abb. 60: Konkordanzen zur Abfrage: Wörter mit (-)*lust*(-) (InterCorp_de)

Query Type:	
CQL	Corpus Query Language: Abfrage nach morphosyntaktischen Zeichen, die jeder Position im Korpus zugewiesen sind. Diese Abfrage ist von ihrer Aufstellung her ein wenig kompliziert, kann aber das Suchen nach vielen sprachlichen Elementen wesentlich erleichtern und Interessantes über die Sprache aufdecken.

Bei Abfragen im CQL-Modus sind immer die Sprache und das dazugehörige Tagset zu beachten! (Viele Sprachen sind im InterCorp noch nicht morphosyntaktisch annotiert - siehe Tab. 15 auf der Seite 80.) Die Verwendung eines „fremden“ Tagsets liefert verständlicherweise keine Ergebnisse – der Korpusmanager „versteht“ die Anfrage nicht.

Die deutschen Korpora im InterCorp (intercorp_de und deWac) sind mit TreeTagger annotiert (genauso wie auch das Archiv TAGGED-T im DeReKo, das DWDS und das CHTK). Das Tagset (Verzeichnis der Zeichen für einzelne morphosyntaktische Kategorien) ist im Kap. 8.2 zu finden. Beispiele für konkrete Abfragen sind im Kap. 8.4 angeführt.

Wenn der Modus CQL eingeschaltet ist, muss die Eingabe immer in der folgenden Syntax (Eckklammern) geschrieben werden. Als Abfragevariable (Attribut) muss word (= Wortform), lemma (= Grundform/ Lemma) oder tag (= morphologisches Zeichen) eingegeben werden.

Eingabe	Ergebnis (Beschreibung)
[word="lustig"]	ergibt nur die Formen *lustig* (ident mit Query Type: Word Form – siehe oben).

Eingabe	Ergebnis (Beschreibung)
[lemma="lustig"]	ergibt alle Flexionsformen des Wortes *lustig*: *lustig, lustige, lustiges, lustigste, Lustigsten* etc. (ident mit Query Type: Lemma – siehe oben).

Eingabe	Ergebnis (Beschreibung)
[tag="ITJ"]	ergibt alle Interjektionen (ITJ = Interjektionen) im Korpus (bzw. alle Wörter, die als Interjektionen erkannt wurden): *na, ach, och, au, hallo, oh, ah, aha, nu, he* sind die häufigsten im InterCorp. (*Au* müsste überprüft werden - siehe letztes Beispiel in den Konkordanzen.)

Die Liste der morphosyntaktischen Zeichen (Codes, die man anstatt ITJ eingeben kann,) ist im Kap. 8.2.

Konkordanzen (Beispiele, Zufallsauswahl):

. Ingrid , lese ich auf dem Display . »	**Hallo**	« , sage ich . » Wie war die Sauftour
Sinn , daß er es zur Kenntnis nehme . »	**Ach**	je « , sagte sie und
„ Jetzt sucht mich ,	**ha**	! Jetzt bin ich es , der besser sieht !
MÖBIUS Ihr besitzt Geheimsender ? EINSTEIN	**Na**	und ?
Erde saß und vergnügt vor sich hin sang : "	**Hurra**	, wie kann ich gut -
berechtigten Grimm . - Au , au , au ,	**au**	! schrie der Kopf
en meine allzu mächtige Portion Mousse	**au**	chocolat an .

Abb. 61: Konkordanzen zur Abfrage: Interjektionen (InterCorp_de)

Kombinationen

Die drei Attribute/ Abfragevariablen [word], [lemma] und [tag] lassen sich auch kombinieren, wenn man aus mehreren homonymen Formen nur eine abrufen will. Aus unzähligen Kombinationsmöglichkeiten werden im Folgenden drei Abfragebeispiele mit Ergebnissen angeführt.

Beispiel 1: Gesucht wird nach der Interjektion (= Code ITJ) *marsch (!)*. Das Substantiv *Marsch* soll ausgeblendet werden.

Eingabe (CQL)	Ergebnis (Beschreibung)
[tag="ITJ"&word="marsch"]	ergibt Interjektionen (ITJ = Interjektionen), die gleichzeitig die Form *marsch* haben.

Konkordanzen (Beispiele, Zufallsauswahl):

, wie Leid dir DAS tun wird . Also ,	**marsch**	in die Wanne , und jetzt (…) kein Wort
Magst du Kaffee ? Katrin ,	**marsch**	ins Zimmer ! "
„ Bißchen nassauern , was ?	**Marsch**	, abfahren ! “ Albert steht wie betäubt

Abb. 62: Konkordanzen zur Abfrage: Interjektion *marsch* (InterCorp_de)

Beispiel 2: Gesucht wird nach dem Wort(teil) *zu*, das nur als Verbzusatz vorkommt.

Eingabe (CQL)	Ergebnis (Beschreibung)
[word="zu"&tag="PTKVZ"]	ergibt Verbzusätze (PTKVZ = Verbzusatz), die gleichzeitig die Form *zu* haben.

Konkordanzen (Beispiele, Zufallsauswahl):

, also steht mir auch ein Drittel von allem	**zu**	. Egal
aber dann drückte er ein Auge	**zu**	, denn
Die Kommission stimmt der Empfehlung	**zu**	.
Sangra hörte wortlos	**zu**	. Aber als der Architekt gegangen war ,
Der Diener ließ das aber nicht	**zu**	. » Nein « , sagte er , » Sie müssen
» Hört ihr da drüben eigentlich	**zu**	? « Professor Raue-Pritsches Stimme
nahm die Hitze jedoch immer mehr	**zu**	, bis sie schließlich kulminierte
und ich stimme	**zu**	, dass die Informationen bezüglich
Dann nickte er mir	**zu**	und ging .
Harry sah eine Weile	**zu**	und stellte fest ,

Abb. 63: Konkordanzen zur Abfrage: Verbzusatz *zu* (InterCorp_de)

Beispiel 3: Gesucht wird nach einer Präposition (= Code APPR) und dem Wort *König* als Gattungsname (= Code NN) im Abstand von max. 2 Wörtern.

Eingabe (CQL)	Ergebnis (Beschreibung)
[tag="APPR"][]{0,2}[lemma="König"&tag="NN"]	ergibt Präposition, *König* als Gattungsname, dazwischen 0-2 andere Wörter.

Konkordanzen (Beispiele, Zufallsauswahl):

Ablehnung der	**von den fränkischen Königen**	praktizierten
wandte sich Judit Kinski	**an den König**	. »
, dachte der Jüngling	**an den alten König**	und den Marktplatz ,
beugte sich Judit	**über den König**	und zog
und wenn es	**um einen König**	längstvergangener Zeiten
als sie auch selbst	**für die Könige**	der Menschen
und wie er	**unter deren legendären Königen**	damals gearbeitet
war nicht persönlich	**mit dem König**	bekannt .
Pluralismus	**unter Königen**	, die Böhmen

Abb. 64: Konkordanzen zur Abfrage: Präposition (2 Positionen) vor dem Appellativum *König* (InterCorp_de)

Vorsicht! Immer auf den Abfragemodus achten. Wird z.B. beim eingeschalteten CQL Modus ein einfaches Wort eingegeben, zeigt das Korpus keine Treffer („Empty result“).

B. Abfragefenster/ Search in the corpus

(Fortsetzung – siehe Abb. 51 auf Seite 82)

Specify context::	
	Beim Anklicken dieser Überschrift öffnet sich das Fenster **Lemmafilter**, in dem der Kontext des Suchbegriffs spezifiziert werden kann. Diese Funktion ist wichtig, wenn die Umgebung des KWIC eingeschränkt werden soll.

Beispiel: Wo kommen im Umfeld von *lustig* auch Lemmata *lachen* und *laut* vor?

Specify context:		
Window:	both = suche rechts und links vom KWIC right = suche nur rechts vom KWIC left = suche nur links vom KWIC 1-15 tokens = Abstand vom KWIC	
Lemma:	Eingabe der gewünschten Lemmata, die im Kontext gesucht werden sollen (hier *lustig* und *lachen*). Gleich daneben: Einstellung des Suchmodus:	
	all	sucht im Umfeld sowohl (*lachen*) als auch (*laut*)
	any of these items	sucht entweder (*lachen*) oder (*laut*)
	none	sucht weder (*lachen*) noch (*laut*) im Umfeld (von *lustig*)

Konkordanzen (alle Belege):

Die Cousine **lachte laut** heraus , weil sie tatsächlich zuviel gekifft hatte und alles als	**lustig**	empfand
und OCCASIONAL LOVER . " Cecilia Vanger **lachte laut** . " Was ist daran so	**lustig**	? "

Abb. 65: Konkordanzen zur Abfrage: Lemma *lustig* im Kontext mit Lemmata *lachen* und *laut*

Specify query according to the meta-information:	
	Die Ergebnisse lassen sich auf bestimmte Texte einschränken. Die Kriterien der Einschränkung sind ident mit der Funktion **Erstellung eines Subkorpus** (siehe dort).

Aligned corpora:	
	Aus dem Angebot unter **corpus** **intercorp_ar** bis **intercorp_uk** (siehe Tab. 13 auf S. 79) wird die gewünschte Parallele durch Klicken auf die entsprechende Sprache und Bestätigen auf den Button Add ausgewählt. **Add** anklicken.

Jetzt erscheint ein Abfrageformular für die erste Parallelsprache. Wiederholt man den Vorgang, erscheint ein Formular für die zweite Parallelsprache (Abb. 66 auf der nächsten Seite).

Es wird also in drei Sprachen recherchiert: Deutsch (als 1. Parallele) – Englisch (als 2. Parallele) – Spanisch (als 3. Parallele).

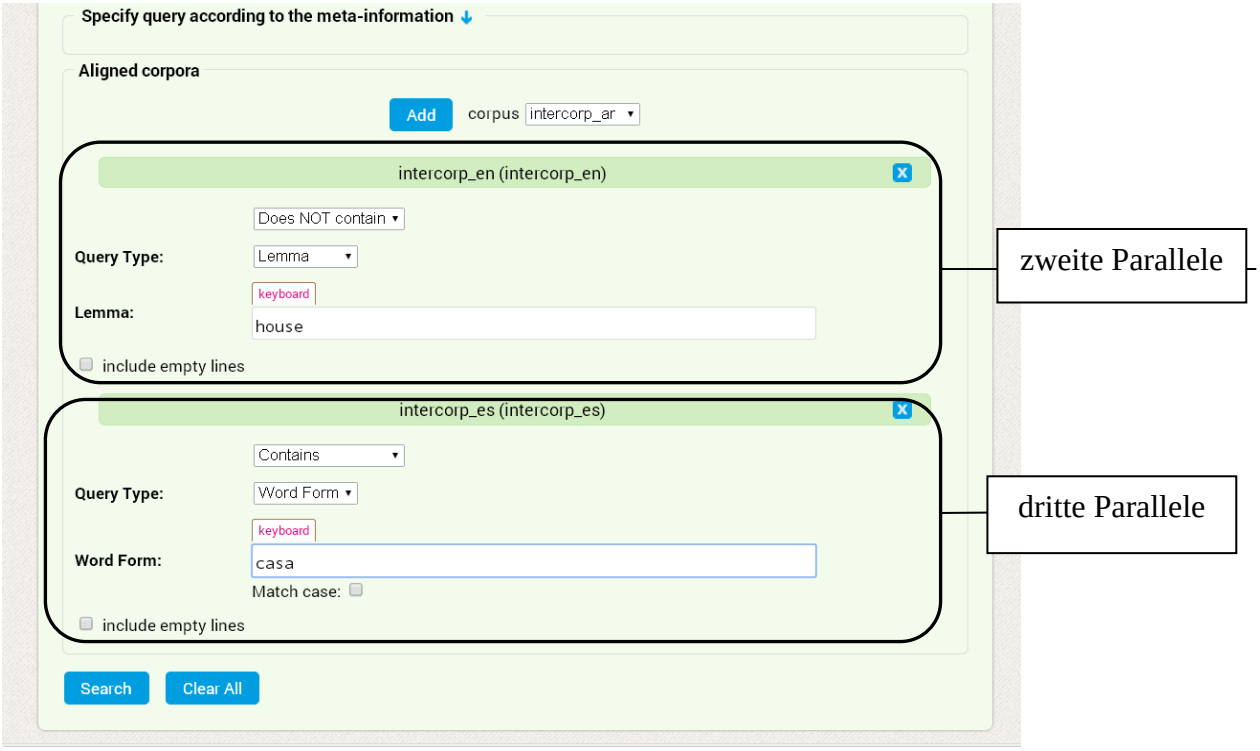

Abb. 66: Aktivierung von drei Parallelen im InterCorp

Jede Parallele kann durch Klicken auf ✖ wieder gelöscht werden.

Wenn kein Formular für die Parallelsprache(n) (zweite, dritte Parallele) ausgefüllt wird, bekommt man als Suchergebnisse alle Belege angezeigt, die der Abfrage in der ersten Sprache/Parallele (hier Deutsch) entsprechen. (Im Beispiel unten sind es alle Segmente, in denen im Deutschen die Wörter *Haus*, *Hause*, *Hauses*, *Häuser* vorkommen.)

Diese Formulare dienen der sog. parallelen Suche, die die Abfrage in der ersten Sprache (hier Deutsch) einschränken: ins Suchfeld der zweiten (dritten …) Sprache kann ein Wort, Lemma, eine Phrase etc. eingegeben werden. Wenn dies erfolgt, sucht das Korpus nur nach denjenigen Segmenten, in denen die Sucheingaben von beiden (allen drei) Sprachen vorkommen (= Auswahl: **Contains**) oder eben nicht vorkommen (= Auswahl: **Does NOT contain**).

include empty lines ☑ (aus-/abwählen)	bedeutet, dass in den Ergebnissen auch leere Segmente erscheinen, also diejenigen Segmente, die in der Parallele der zweiten (dritten …) Sprache keine Entsprechung haben. Dies kann mehrere Ursachen haben: a) translatologische Arbeit (Übersetzungen entsprechen einander kaum); b) andere Auflage der Publikation; c) schlechtes Alignment.

Beispiel:
Gesucht wird nach Entsprechungen des deutschen Wortes *Schule* im Englischen und Spanischen, wobei man im Englischen das Wort *school* ausblenden möchte.

Search in the corpus

Corpus:	intercorp_de	auswählen
Query Type:	Basic	auswählen
Query:	**Schule**	eingeben

Aligned corpora

Corpus:	intercorp_en	auswählen→ **Add**
	Does NOT contain	auswählen
Query Type:	Lemma	auswählen
Lemma:	**school**	eingeben

dann

Corpus:	intercorp_es	auswählen→ **Add**
	Contains	auswählen
Query Type:	Word form	auswählen
Word form:	**escuela**	eingeben
		Search anklicken

Konkordanz (Beispiel, Zufallsauswahl):

de (links)	KWIC	de (rechts)	en	es (links)	KWIC	es (r.)
Man hätte mich von der	**Schule**	geworfen .	I 'd have been kicked out of the Academy .	Me hubieran echado de la	**escuela**	.

Abb. 67: Konkordanz zur Abfrage: (de) Basic: *Schule* - (en) NICHT Lemma *school* – (es) Wortform *escuela* (InterCorp_de-en-es)

Soll im Spanischen auch das Wort *escuela* ausgeblendet werden, stellt man einfach den Schalter Contains auf Does NOT contain um.

Konkordanz (Beispiel, Zufallsauswahl):

de (links)	KWIC	de (rechts)	en	es
Er war aus der Partei und aus der	**Schule**	hinausgeflogen .	He 'd been kicked out of both the Party and the university .	Lo habían echado del partido y de la facultad .

Abb. 68: Konkordanz zur Abfrage: (de) Basic: *Schule* - (en) NICHT Lemma *school* – (es) NICHT Wortform *escuela* (InterCorp_de-en-es)

Die Konkordanzen/ Segmente erscheinen in Spalten und lassen sich einfach exportieren (siehe unter Export).

C. Einstellungen und Funktionen

Auf der oberen Leiste (neben dem Logo KonText) findet man (waagrecht gereiht) folgende Posten:
Query|Subcorpora|Save|Concordance|Filter|Frequency|Collocations|View options|Help

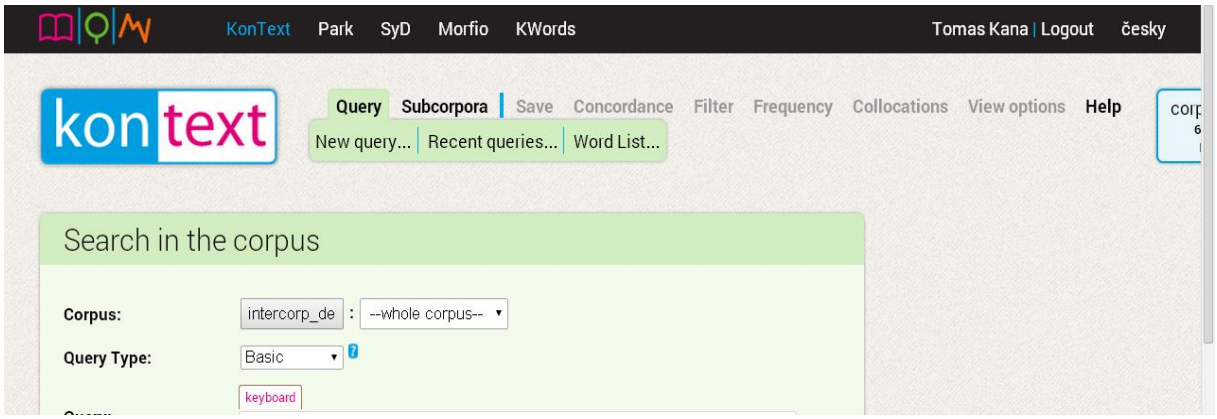

Abb. 69: Einstellungen und Funktionen: Level 1 (InterCorp)

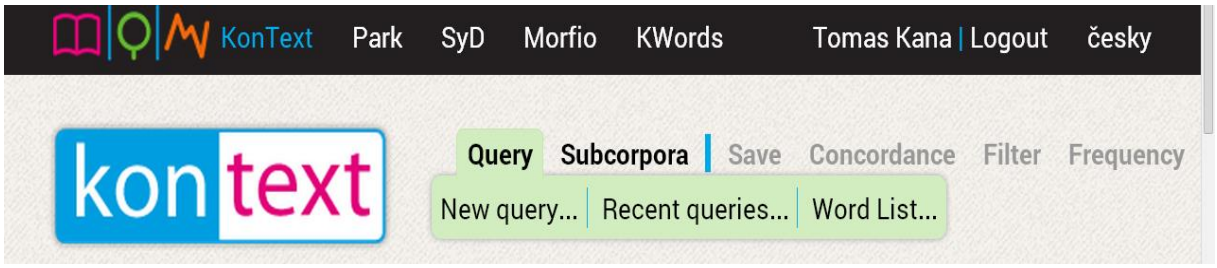

Abb. 70: Einstellungen und Funktionen: Level 2 (InterCorp)

Fährt man mit der Maus zu dem gewünschten Posten, öffnet sich unter ihm das weitere Angebot (hier unter Query: New query | Recent quiries | Word List).
Die Posten Save bis View Options sind erst nach dem Aufrufen der Konkordanz aktiv.

Query

New query…|Recent queries…|Word List…

Hier gelangt man zu den Abfragefenstern.

New query	neue Anfrage (siehe **B. Abfragefenster/Search in the corpus** ab Seite 84.)
Recent queries	zuletzt gestellte Anfragen während der letzten Sitzungen. Diese können wieder abgerufen werden.
Word List	gibt die Frequenzliste der Wörter im aktuellen Korpus an. Die Liste erscheint automatisch beginnend mit dem Wort mit der höchsten Frequenz, lässt sich aber auch alphabetisch erstellen (durch Klicken auf Word).

Subcorpora

Create new|My subcorpora…

Diese Funktion ermöglicht die Einschränkung/ systematische Auswahl aus dem bestehenden Korpus. Wichtig ist dies für eine gezielte Suche in spezifischen Texten. Das neue Subkorpus kann man nach verschiedenen Kriterien (siehe unten) aufbauen. (Die Kriterien sind gleich wie bei der Einschränkung der Konkordanzen bzw. beim Spezifizieren der Suchkriterien – siehe Specify query according to the meta-information auf Seite 91.)

Create new	
New subcorpus name:	Benennen des neuen Subkorpus: einen kurzen, treffenden Namen eingeben.
Specify subcorpus using: Custom 'within' condition ⊙ Attribute list ⭘	spezifiziert die Subkorpuscharakteristik aus-/abwählen
Custom 'within' condition Attribute list ⊙	erfordert eine gute Orientierung in den Korpustexten und in ihren Bezeichnungen. Die Erstellung nach diesen Kriterien ist etwas „benutzerunfreundlich“. Einfacher ist ein Vorgehen über den Befehl Attribute list. Auswahlmöglichkeiten über Aus-/Abwahlbuttons ☑

div.group ☑ jádro/ core ☑ kolekce/ collections	Das „**Kernkorpus**“ (jádro/ core) besteht aus manuell bearbeiteten Texten. Die „Textsammlungen“ (kolekce/ collections) bestehen aus automatisch bearbeiteten Internettexten.
div.id: ☑ _ACQUIS ☑ _EUROPARL …	**Technische Namen**[20] einzelner Texte im Korpus.
div.txtype: ☑ drama ☑ literatura faktu …	**Stilistische Charakteristika der Texte** Da diese Werte derzeit nur auf Tschechisch erscheinen, gibt es auf der nächsten Seite eine kleine sprachliche Hilfe.
div.srclang: ☑ - (unspezifisch) ☑ bg. ☑ cs …	**„Source language“ – Wahl der Originalsprache**
div.original: ☑ - (unspezifisch) ☑ ano/ yes ☑ ne/ no	**Originaltexte** Wählen Sie **ano**, wenn sie nur in (z.B. deutschen) Originaltexten browsen wollen.

[20] Diese werden im Laufe der Zeit durch leichter identifizierbare Titel ersetzt.

Übersetzungshilfe:

Attribute (siehe vorherige Seite)	**Tschechisch**	**Deutsch**
div.group	jádro/ core kolekce/ collections	Kernkorpus (manuell alignierte Texte) Sammlungen (automatisch alignierte Texte)
div.txtype	drama literatura faktu poezie právní texty próza publicistika - komentáře publicistika - zprávy různé zápis debaty	dramatische Texte Sachtexte Poesie rechtswissenschaftliche Texte Prosatexte publizistische Texte – Kommentare publ. Texte – Nachrichten u. Berichte Diverses Gespräche
div.original:	ano/ yes ne/ no	ja nein

Save

CSV|XML|TXT|Custom…

Hier werden die Recherchergebnisse gespeichert. Sowohl Konkordanzen als auch Frequenzlisten lassen sich in verschiedenen Formaten speichern (CSV, XML, TXT): man muss mit der Maus auf die Ikone **Save** gehen, dann **Custom** wählen. Es eröffnet sich eine Abfragetabelle:

Custom…	
Save concordance as:	Speicherformat einstellen: Text (txt), XML oder CSV. Empfehlenswert ist das CSV-Format (einfache Konversion in eine Excel-Tabelle), denn die KWICs werden in einer separaten Spalte zwischen dem Kontext links und rechts gespeichert. (Danach folgt die Arbeit mit Excel: Indexierung, Sortieren etc.)
Include heading:	aus-/abwählen (nur bei der Wahl vom TXT- oder XML-Format). Mitgespeichert werden dabei Informationen über die Abfrage: Korpus, Eingabe, Abfragemodus etc.
Include line numbers:	Falls ausgewählt, werden die Konkordanzen nummeriert.
Align KWIC:	Falls ausgewählt, sind die KWICs in der Zeilenmitte und untereinander geordnet.
Lines to store:	In diesem Fensterchen kann die Anzahl der zu speichernden Konkordanzen eingestellt werden.

Concordance

Current concordance|Sorting…|Shuffle|Sample…|Query overview…|Undo

Arbeit mit Konkordanzen: sortieren, mischen, Auswahl der Treffer, Rückkehr zu vorherigen Schritten.

Current concordance	Schritt zurück zu der zuletzt abgerufenen Konkordanzliste
Sorting…	sortiert Konkordanzen nach mehreren Parametern: **Simple Sort:** sortiert Konkordanzen nach dem KWIC oder dessen Kontext rechts oder links (Sort Key). Die Sortierung erfolgt alphabetisch (automatisch) oder retrograd (Backward). Sortierungskriterien sind „Attribute". Als „Attribute" gelten: word, lemma, tag (falls getagget) (siehe **Query Type)**, weiterhin Name des Dokuments, Geschlecht des Autors etc. **Multilevel sort:** ermöglicht die Sortierung nach Kombinationen von mehreren Kriterien.
Shuffle	mischt die Konkordanzzeilen erneut.
Sample	Nach dem Anklicken kann die Anzahl der Belege reduziert werden (Zufallsauswahl).
Query overview…	Übersicht über die einzelnen Schritte bei der letzten Abfrage. So gelangt man schnell zu den Ergebnissen, z.B. vor einer durchgeführten Sortierung (durch Klicken auf View result).
Undo	ein Schritt zurück.

Filter

Positive…|Negative…

Filtern der Konkordanzen: Der Filter ermöglicht, die Belege nach der Umgebung von KWIC zu sortieren. Mit dieser Funktion kann nach sprachlichen Verbindungen gesucht werden. Nach dem Klicken auf **Positive...** oder **Negative...** öffnet sich dasselbe Formular:

Positive…|Negative…	
Concordance Filter	
positive:	zeigt Belege, die den angegebenen Bedingungen entsprechen.
negative:	löscht Belege mit der angegebenen Bedingung.
Selected token:	Wahlmöglichkeit bei mehrfachem Vorkommen (erstes oder letztes Vorkommen) desselben Elements im definierten Kontext (Span). Diese Funktion ist nur für positives Filtern relevant. **Beispiel:** Abfrage Lemma **reden** und Suche nach **zu** in seinem Kontext: **first:** Lukas schien auch keine Lust **zu** haben , viel zu **reden**. **last**: Lukas schien auch keine Lust zu haben , viel **zu reden**.
Search Span:	Einstellung der Spannbreite des Suchkontextes (Abstand links vom KWIC links: negative Ziffern; rechts vom KWIC: positive Ziffern)
Query Type:	Basic, Lemma, Phrase etc. (siehe B. Abfragefenster)
Query:	Sucheingabe (siehe B. Abfragefenster)

Frequency

Node tags|Node forms|Doc IDs|Text Types|Custom…

Informationen über die Häufigkeiten, Abrufen von Frequenzlisten mit statistischen Angaben über KWICs (Node).

Node tags
Nach dem Anklicken bekommt man die Häufigkeitsliste von Tags der KWICs zur aktuellen Abfrage.

Beispiel: abgefragt wird die Form *runde*.
Abfrage: Basic: **runde** → Node tags (Abrufen der Frequenzliste – Abb. 71)

Frequenzliste:

		tag	Freq	Freq [%]
1.	p/ n	NN	1,075	78.6
2.	p/ n	ADJA	280	20.5
3.	p/ n	VVFIN	12	0.9

Abb. 71: Frequenzliste der Tags zur Abfrage *runde* (Basic)

Über diesen Weg erfährt man schnell, ob das Substantiv (*die Runde*) oder das Adjektiv (*runde*) im Korpus überwiegen.
Aus der Abb. 71 ist ersichtlich, dass etwa 88% auf das Substantiv *Runde* und 10% auf das Adjektiv *runde* entfallen. Nur 2% bildet das Verb *abrunden* mit abgetrenntem *ab-* .

Node forms
Bei einer Abfrage, die mehr als ein Ergebnis liefert: Lemma, Word Form etc. bekommt man nach dem Anklicken die statistische Übersicht über die einzelnen Formen der Ergebnisse.

Beispiel: abgefragt wird die Grundform des Verbs *fragen*.
Abfrage: Lemma: **fragen** → Node forms (Abrufen der Frequenzliste)

Frequenzliste[21]**:**

		word	Freq	Freq [%]
1.	p/ n	fragte	13,646	57.4
2.	p/ n	fragen	3,881	16.3
3.	p/ n	gefragt	1,961	8.2
4.	p/ n	fragt	1,859	7.8

Abb. 72: Frequenzliste der Formen zur Abfrage *fragen* (Lemma)

Diese Funktion ist sehr wichtig für die Erstellung von Frequenzlisten einzelner sprachlicher Erscheinungen (mehr dazu in den Fallbeispielen).

[21] In publizistischen Texten kommen die Formen in dieser Reihenfolge der Frequenz vor: *fragen, frage, gefragt, fragt, fragte, fragten*; in belletristischen Texten: *fragte, fragen, fragten, fragst, gefragt*.

Doc IDs
Graphik der Aufteilung von Belegen in einzelnen Dokumenten

Text Types
Graphik der Aufteilung von Belegen in einzelnen Texttypen: sortiert nach Korpus, Dokument, Stil des Textes, Originalsprache etc.

Custom…		
Benutzerdefinierte Einstellung der Frequenzliste Diese Applikation ermöglicht die Sortierung der Konkordanzen nach mehreren Kriterien und auf mehreren Ebenen (Level). Als Variablen der Sortierungskriterien gelten das Attribut des zu berücksichtigenden Elements und der Abstand zum KWIC.		
Level	Sortierungsebene	
Attribute	Auswahl der Charakteristik (Attribut) der Elemente in der Umgebung	
	word	Wortform (Groß-/Kleinschreibung beachten)
	lemma	Lemma (Groß-/Kleinschreibung beachten)
	tag	Tag
	lc	Wortform (Groß-/Kleinschreibung ignorieren)
	lemma_lc	Lemma (Groß-/Kleinschreibung ignorieren)
Ignore case?	☑	Groß-/Kleinschreibung beachten/ ignorieren
Position	**6L** **Node** **6R**	Abstand (vom KWIC) derjenigen Elemente, die in die Frequenzliste einbezogen werden sollen: von (max.) 6 Positionen links bis (max.) 6 Positionen rechts vom KWIC
(Node) start at	**leftmost KWIC word** **rightmost KWIC word**	Berechnung ab dem linken/ rechten Ende eines mehrgliedrigen KWICs

Collocations

Custom…

Diese Funktion (nach dem Anklicken von Custom…) ermöglicht die automatische Auflistung der häufigsten Kollokationspartner/ Kollokatoren zu KWICs (entspricht der Kookkurrenzanalyse im DeReKo).
Die Berechnung der Kollokationen deckt lockere Verbindungen, die häufig vorkommen, aber auch Phraseme (=feste Verbindungen) auf.

Die Berechnung und Aufstellung von Listen der Kollokationspartner erfolgt nach komplizierten statistischen Formeln. Als Kollokationspartner können Wörter, Lemmata oder Tags berechnet werden.
Beim Anklicken von Collocations → Custom… erscheint das Abfragefenster
Collocation candidates, in dem folgende Posten einzustellen sind:

Attribute:	Auswahl der Charakteristik (Attribut) des Kollokationspartners:	
	word	Wortform (Groß-/Kleinschreibung beachten)
	lemma	Lemma (Groß-/Kleinschreibung beachten)
	tag	Tag
	lc	Wortform (Groß-/Kleinschreibung ignorieren)
	lemma_lc	Lemma (Groß-/Kleinschreibung ignorieren)
In the range from	-5 **to:** 5	Abstand vom KWIC: links vom KWIC: negative Ziffern (Default -5) rechts vom KWIC: positive Ziffern (Default 5)
Minimum frequency in corpus 5		In die Liste kommen keine Kollokationspartner, die im Korpus seltener vorkommen als die angegebene Ziffer (Default 5).
Minimum frequency in given range 3		In die Liste kommen keine Kollokationspartner, die in der Umgebung vom KWIC seltener vorkommen als die angegebene Ziffer (Default 3).
Show functions:	Auswahl der Signifikanzmaße	
	T-score	☑
	MI	☑
	MI3	☑
	log likelihood	☑
	logDice	☑ …
Sort by:	Auswahl der Sortierung	
	T-score	O
	MI	O
	MI3	O
	log likelihood	⊙
	logDice	O …

Beispiel: Gesucht wird nach Kollokationen zum Verb *fragen*.
Korpus: Intercorp_de → InterCorp_en, Version 6 (2013)

Query Type:	Lemma	auswählen
Lemma:	**fragen**	eingeben
		Search anklicken

Nachdem die Konkordanzen erscheinen:

Collocations → Custom…		auswählen
Collocation candidates	Formular folgend ausfüllen:	
Attribute	lemma	auswählen
Range	from -5 to 5	eingeben
Minimum frequency in corpus	100	eingeben
Minimum frequency in given range	15	eingeben
Show functions	log likelihood logDice	auswählen
Sort by:	log likelihood	
		Make candidate list anklicken

Als Ergebnis erscheint eine Tabelle mit Kollokationskandidaten. (Bemerkenswert ist die Geschwindigkeit, in der die Kandidaten errechnet und die Tabelle erstellt werden.) Die Reihung erfolgt immer nach dem angegebenen Wert absteigend (Abb. 73). Durch das Klicken auf einen anderen Wert (falls seine Berechnung verlangt wurde) reihen sich die Kandidaten nach diesem neu. (Abb. 74).

Total: 884					
	Lemma		**Freq**	**log likelihood**	**logDice**
1.	p/n	**?**	10714	101988.599	11.128
2.	p/n	**.**	13560	59077.338	7.512
3.	p/n	**,**	14295	49574.687	6.819
4.	p/n	**ich**	6868	38689.910	8.682
5.	p/n	**«**	5084	38402.586	9.973
6.	p/n	**er**	6194	35647.718	8.783
7.	p/n	**»**	3928	26931.233	9.524
8.	p/n	**ob**	2879	26884.599	10.723
9.	p/n	**"**	3923	19828.580	8.334
10.	p/n	**sie**	3462	16962.968	8.230
11.	p/n	**was**	2418	15984.957	9.304

Abb. 73: Kollokatoren zum Lemma *fragen*: Lemmata, sortiert nach log likelihood (InterCorp_de)

Total: 884					
	Wort		**Freq**	**log likelihood**	**logDice**
1.	p/n	**?**	10714	101988.599	11.128
2.	p/n	**ob**	2879	26884.599	10.723
3.	p/n	**«**	5084	38402.586	9.973
4.	p/n	**warum**	1042	9092.661	9.783
5.	p/n	**“**	2100	15011.059	9.566
6.	p/n	**ihn**	3928	26931.233	9.524
7.	p/n	**was**	2418	15984.957	9.304
8.	p/n	**du**	2185	14425.423	9.276
9.	p/n	**mich**	1663	11085.375	9.232
10.	p/n	**wer**	625	4813.691	9.056
11.	p/n	**danach**	904	5792.607	8.844

Abb. 74: Kollokatoren zum Lemma *fragen*: Lemmata, sortiert nach logDice (InterCorp_de)

Diakritische Satzzeichen (Fragezeichen, Anführungszeichen, Punkt, Beistrich etc.) kommen sehr häufig vor, genauso wie Subjekte *er*, *ich* und Objekte *ihn*, *mich* (Abb. 74), sowie die typischen Rektionsoperatoren (Präposition *nach* oder einleitende Konjunktion zum Objektsatz *ob*).

Erklärungen zu den Abb. 73 und 74:

Total: 884	Anzahl der errechneten Kollokatoren (hier 884)
p/n	Positiv/ negativ filtern: Durch Klicken auf p gelangt man zu Konkordanzen mit dem konkreten Kollokationspartner. Zu *ob* (8. Kollokationspartner) siehe Abb. 75. Durch Klicken auf n bekommt man erneut alle Belege der Abfrage, nur diejenigen mit dem konkreten Kollokationspartner erscheinen nicht. (Ohne *ob* siehe Abb. 76).
Lemma **Word**	Attribut, nach dem die Kollokationspartner gesucht wurden. Es werden also nur die Grundformen angezeigt, z.B. hier unter *ich* (4. Kollokationspartner) wurden auch die Formen *meiner*[22], *mir* und *mich* gerechnet. Will man konkrete Formen abrufen, muss das Attribut im Formular Collocation candidates auf word umgestellt werden (Ergebnisse siehe Abb. 74.)
Freq	Absolute Frequenz des gemeinsamen Vorkommens.
log likelihood	Signifikanzmaße (siehe S. 17-18)
logDice	

[22] Diese Form allerdings nur theoretisch. Im Korpus gibt es keinen Beleg des Genitivs von *ich*.

Beispiele:

Konkordanzen zum positiven Filter zu *ob* (8. Kollokationspartner in der Abb. 73); Zufallsauswahl aus 2.879 Belegen (siehe Spalte Freq.):

Darf ich mir erlauben , Sie zu	**fragen**	, ***ob*** Sie als Kind allein aufwuchsen
Sie	**fragte**	, ***ob*** sie mal mein Tagebuch lesen dürfte .
Wolfgang Pauli wurde einmal	**gefragt**	, ***ob*** eine … Abhandlung … falsch sei .
Er öffnete die Tür zum Nebenraum und	**fragte**	Jaynes , ***ob*** er eine Minute Zeit hätte

Abb. 75: Konkordanzen zur Abfrage: Lemma *fragen … ob* (InterCorp_de)

Konkordanzen zum negativen Filter zu *ob* (8. Kollokationspartner in der Abb. 73); Zufallsauswahl aus 6.733 Belegen (Gesamtanzahl der Belege Lemma *fragen* = 8.271, ausgeblendet werden alle Belege, in denen *ob* vorkommt (2.879)):

und bei der man sich heute	**fragt**	, wie viele Leute … der … Meinung sind .
… und ich	**frage**	Sie in aller Deutlichkeit , …
» (…) War mir egal . Ich hab ihn nicht	**gefragt**	« »Wie lange haben Sie dort gewohnt?«

Abb. 76: Konkordanzen zur Abfrage: Lemma *fragen …* ohne *ob* (InterCorp_de)

Konkordanzen zum positiven Filter *danach* (11. Kollokationspartner in der Abb. 74), Zufallsauswahl.

Warum sollte man ihn nicht **danach**	**fragen**	? Er konnte schließlich antworten ,
Niemand	**fragte**	mich **danach** .
. Ich habe nicht **danach**	**gefragt**	, und seinen Worten habe …

Abb. 77: Konkordanzen zur Abfrage: Lemma *fragen … danach* (InterCorp_de)

Über den positiven Filter p werden also Konkordanzen aufgerufen, aus denen Chunks mit dem Verb *fragen* abgeleitet werden können. Die automatische Erstellung von syntagmatischen Mustern, wie es COSMAS II (S. 69) ermöglicht, bietet KonText nicht an.

View options

KWIC/Sentence|Attributes, structures and references…|General concordance view options…

Hier lässt sich die Ansicht der Konkordanzen nach unterschiedlichen Kriterien einstellen.

KWIC/Sentence:
Umschalten zwischen der Ansicht der Konkordanzen in Form von KWICs (die KWICs in der Mitte und Konkordanzen zentriert) oder Sätzen (die Konkordanzen sind nach rechts gerückt und umfassen die Länge eines Satzes).

Attributes, structures and references	
Attributes	Anzeigen von Attributen einzelner Tokens in den Konkordanzzeilen ☑ word ☑ lemma ☑ tag …
O Display attributes for each token:	Attribute werden bei jedem Token angezeigt.
⊙ Display attributes for KWIC tokens only:	Attribute werden nur bei den KWICs angezeigt.

Bemerkung: Falls die Option Display attributes for each token gewählt wird, ist der Text in der Konkordanzzeile erschwert lesbar.
Diese Funktion ist sehr hilfreich, wenn man folglich nach Lemmata und/ oder Tags suchen möchte, die in der Umgebung vom KWIC vorkommen, man weiß aber nicht, wie das Tagset systematisiert ist.

Structures	Anzeigen von (HTML-)Strukturzeichen in den Konkordanzzeilen: Ende des Satzes, Absatzes etc. ☑ <opus> ☑ <doc> ☑ <s>
References	Metainformationen zum Text, in dem die Konkordanz vorkommt. ☑ token number ☑ Document number ☑ opus.autor…

General concordance view options		
Result range and paging		
Page size (number of lines):	40	Anzahl der Belege auf einer Seite (Default 40)
KWIC Context size (positions):	10	Anzahl der Positionen in den Konkordanzen vom KWIC rechts und links (Default 10) Empfohlen sind nicht mehr als 20 Positionen. Einen breiteren Kontext (Absatz) kann man immer durch Klicken auf KWIC oder auf ein Wort im parallelen Segment abrufen.
Sort concordance according to good dictionary examples:	☑	(Aus-/Abwahlbutton) Wenn ausgewählt, werden die Konkordanzen nach kurzen, treffenden und abwechslungsreichen Beispielen sortiert. Diese sind dann in der Liste oben angezeigt[23].
Number of lines to be sorted:	100	Anzahl der „guten Beispiele“ (GDEX) (Default 100). Diese erscheinen als erste in der Konkordanzliste (also die ersten 100).
Other options		
Shuffle concordance lines by default	☑	(Aus-/Abwahlbutton) Wenn ausgewählt: bei der neuen Abfrage werden die Konkordanzzeilen gemischt angezeigt. Wenn nicht ausgewählt, werden die Konkordanzen nach einzelnen Dokumenten angezeigt: alphabetisch nach den (technischen) Namen der Dokumenten gereiht.

Help

User manual…|Linguistic terms…|Available corpora…

Das Helpdesk ist derzeit nur in tschechischer Sprache zugänglich, die englische Version ist in Vorbereitung. Eine deutsche Version ist nicht geplant.
Unter Help gelangt man zum User manual…, zu den Definitionen der häufigsten korpuslinguistischen Begriffe (Linguistic terms…) und zur Übersicht über einzelne Korpora

[23] Diese Funktion (abgekürzt GDEX) ist für Lexikograph/-innen gedacht (Kilgarriff et al. 2008), kann jedoch auch beim Erstellen von Lehr- und Übungsmaterialien sehr gut genutzt werden.

und ihre Eigenschaften, welche das Institut vom Tschechischen Nationalkorpus zur Verfügung stellt (Available corpora…)
Das Layout des Helpdesk ist benutzerfreundlich, u.a. auch deswegen, weil es sich an die Form von Wikipedia anlehnt.

5. Konkordanzprogramme und korpusähnliche Instrumente

Konkordanzprogramme (concordancer) sind Softwaretools, die in einem Text oder in einer Textsammlung (Korpus) nach unterschiedlichen sprachlichen Elementen suchen können und diese in Form von Konkordanzzeilen anzeigen. Sie existieren auch unabhängig von Korpora, viele von ihnen werden als Free- oder Shareware ins Internet gestellt.

Korpusähnliche Instrumente sind Software- oder Webtools, die einige Korpuseigenschaften aufweisen, sie können jedoch nicht als Sprachkorpora bezeichnet werden, denn ihnen fehlt eine substanzielle Korpuseigenschaft: entweder basieren sie auf einer zu kleinen Textsammlung oder sie ermöglichen keine Suche nach sprachlichen Elementen, bzw. können keine statistischen Angaben liefern.

5.1. TextSTAT

Aus der Menge der im Internet zugänglichen Konkordanzprogramme, in der sich jede/-r Interessent/-in das Passende für sich aussuchen kann, wird hier ein simples Programm, das keine besonderen Computer-Kenntnisse verlangt, sehr benutzerfreundlich und noch dazu kostenlos ist, vorgestellt: TextSTAT von Mathias Hüning (2014) an der Nederlandistik der FU Berlin. Mit diesem Programm kann man simple Recherchen in beliebigen Texten durchführen. Die Texte werden einfach ohne jegliche vorherige Aufbereitung ins Programm geladen, während für andere Konkordanzprogramme oder Korpora jeder Text von Bildern, Tabellen befreit und konvertiert werden muss. Auch dieser Vorteil macht TextSTAT zum guten Einstiegsinstrument in die Korpuslinguistik.

Arbeit mit TextSTAT[24]
Es gibt kein ausführliches Manual zu diesem Programm im Internet. Basisinformationen und zwei Quicktours auf Englisch sind von der Seite des TextSTAT aus zugänglich.

TextSTAT-Korpuserstellung:
1. Internetseite http://neon.niederlandistik.fu-berlin.de/textstat/ besuchen
2. Programm TextSTAT auf die Festplatte herunterladen
3. Programm öffnen – das Arbeitsfeld erscheint (Abb. 78)

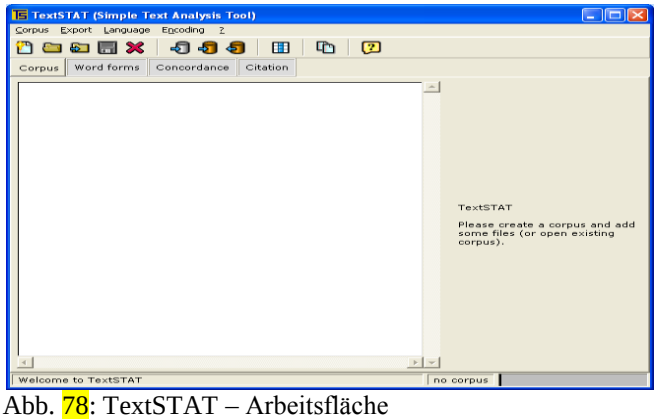

Abb. 78: TextSTAT – Arbeitsfläche

[24] Alle TextSTAT-Screenshots mit freundlicher schriftlicher Genehmigung von M. Hüning.

4. Arbeitssprache einstellen: **Language** (die letzte Version (2.9) „spricht“ Englisch, Deutsch, Niederländisch, Portugiesisch, Spanisch, Katalanisch, Galizisch, Französisch, Italienisch, Finnisch (Suomi), Polnisch, Tschechisch).
5. Korpus anlegen: **Korpus** → **Neues Korpus** oder auf die erste Ikone klicken (Abb. 79)

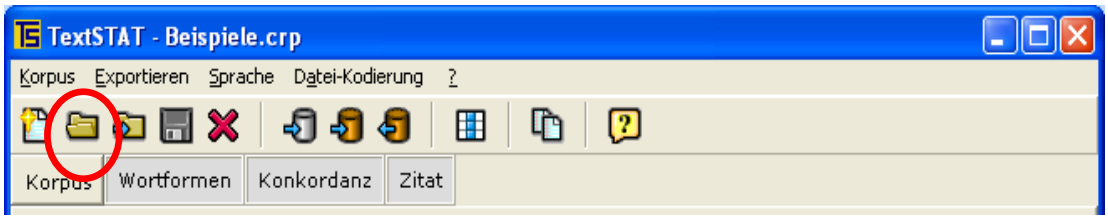

Abb. 79: TextSTAT – Korpus erstellen

6. Datei(en) hinzufügen: **Korpus** → **Datei von Festplatte hinzufügen** oder **Web-Datei hinzufügen** oder die „Walzen-Ikonen“ verwenden (Abb. 80)

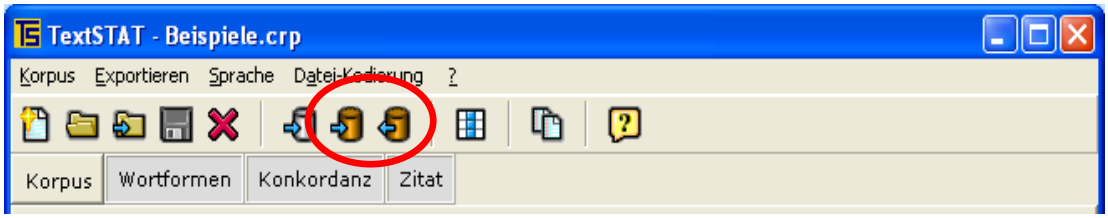

Abb. 80: TextSTAT – Datei laden

Nachdem das Korpus erstellt ist, beginnt die Arbeit mit den sprachlichen Daten.

Abfragen:
Zugang über die Taste Konkordanz

Suchmöglichkeiten im TextSTAT:

Wortform: **Suchfeldeingabe:** (gewünschte Wortform) → Suchen
Wortteil: Suche nach dem Wortanfang **Suchfeldeingabe: \Wxxxxx** → Suchen Suche nach dem Wortende **Suchfeldeingabe: yyyyy\W** → Suchen

Beispiele:

Eingabe	Ergebnis
Mannschaft	*Mannschaft* (nur diese Form)
\Whaupt	*hauptsächlich, Hauptbahnhof* (Wörter mit *haupt-*)
schaft\W	*Bekanntschaft, Leidenschaft, Gastwirtschaft* (Wörter mit *-schaft*)

Erklärungen und Bemerkungen:
Das große **W** steht für ein Leer- oder Satzzeichen, der umgekehrte Schrägstrich hebt die Funktion des Buchstabens *W* auf.
Lemmasuche (suche nach dem Grundwort) oder nach grammatikalischen Kategorien sind im TextSTAT nicht möglich.

Funktionen:

Statistiken:	Unter der Taste Wortformen werden folgende Angaben abgerufen: **Häufigkeit der Wortformen im Korpus** **Liste der Wortformen:** **alphabetisch** **retrograd** **Wörter mit dem String:** Frequenz einer vom User definierten Buchstabenkombination Frequenzliste anklicken
Speichern:	Zugang über die Taste Exportieren **Frequenzliste**: Format xls (csv) wird auf die Festplate gespeichert. **Konkordanzen**: Format doc oder txt werden auf die Festplatte gespeichert.

Einsatz:
Mit diesem Programm kann man auch schnell (unnötige) Wiederholungen oder begriffliche Inkonsequenzen feststellen. Die Listen können nicht nur nach Frequenz, sondern auch alphabetisch oder retrograd (rückläufig) sortiert werden. Darüber hinaus ist auch die Abfrage nach Wortteilen („Wörter mit dem String“) möglich.

Das folgende Beispiel der Arbeit mit TextSTAT soll eine Regularität in der Frequenz der Wörter unabhängig von der Länge und vom Stil eines Textes demonstrieren. Die Versuchstexte wurden aus verschiedenen Sprachbereichen und unterschiedlichen stilistischen Niveaus ausgewählt. Beide Romane sind vergleichbar lang, die beiden Sachtexte wesentlich kürzer. Dem statistischen Vergleich wurden folgende Texte unterzogen:

1) Roman 1: Kundera, Milan (1970): *Scherz*. Ein Roman (113.374 Wörter).
2) Roman 2: Uri, Helene (2008): *Nur die Stärksten überleben*. Ein Roman (100.069 Wörter).
3) Interview in einer Zeitschrift: profil (2009): *Interview: "Hausbesetzer sind konservativ": Blixa Bargeld im Interview mit profil* (1.207 Wörter).
4) Fachartikel (Kapitel aus einem Fachbuch): Sorger, Brigitte et al. (2013): *Schreiben in mehreren Sprachen*. Kapitel Beschreibung einer Handlung (8.249 Wörter).

Jeder Text wurde ins TextSTAT geladen und nach den am häufigsten vorkommenden Wortformen abgefragt. Die Ergebnisse sind in der Tab. 19 zu sehen. Die Ziffern stellen das absolute Vorkommen der Wörter im jeweiligen Text dar.

	Roman 1		Roman 2		Interview		Fachartikel	
1	und	4033	und	3363	ich	31	der	275
2	ich	3783	die	2738	die	27	die	191
3	die	2736	er	2410	das	24	in	189
4	sie	2518	sie	2352	in	21	und	146
5	der	2488	der	1791	ein	20	den	91
6	in	1603	in	1480	nicht	20	auch	63
7	daß	1503	das	1362	Sie/sie	20	sind	58
8	das	1475	zu	1205	ist	18	im	57
9	zu	1411	sich	1169	der	16	ist	57
10	nicht	1344	ist	1156	und	15	auf	56

Tab. 19: Frequenz der Wortformen in vier Texten (TextSTAT)

Auf den ersten Blick sieht man, dass die Listen der zehn häufigsten Wörter in allen drei Texten fast ident sind. Die Differenzen weisen aber bereits auf typische Merkmale der einzelnen Textsorten hin, die in diese Punkte zusammengefast werden können:

- *ich* ist das häufigste Wort in einem Interview sowie in einem in der Ich-Form verfassten Roman (Roman 1);
- Die Wörter *er* und *sie* sind stark frequentiert in einem in der Er-Form verfassten Roman (Roman 2).
- Zu bemerken ist die Absenz des Personalpronomens *ich* unter den ersten zehn häufigsten Wörtern im Fachartikel. (Der Autor bleibt im „Hintergrund".)
- Zur Semantik: alle Wörter repräsentieren inhaltlich „leeres" Wortgut (sich wiederholende Strukturwörter): Deixeis (hier Pronomina), Funktionswörter (Hilfsverb, gram. Partikel), andere Synsemantika (Präpositionen, Konjunktionen), jedoch keine Autosemantika. Dieses Phänomen ist fast in jedem ausformulierten Text zu beobachten. Zum Vergleich: die häufigsten deutschen Wörter in geschriebenen Texten sind: *der/die, und, in, von, mit, zu, das/den, im, für, sich* (DeReWo 2009).
- Je tiefer man in der Liste gehen würde, desto deutlicher kämen die Konturen der Informationssäulen des Textes (Hauptträger der Informationen auf der Wortebene) zum Vorschein: z.B. im Roman 2 kommen etwa ab der zwanzigsten Stelle (Frequenz um 300 Vorkommen) Namen der Protagonisten vor: *Pål, Nanna, Rinkel, Mutter*; weiterhin die Schauplätze *Institut, Büro, Universität* und „Objekte" des Roman-Plots: *Sprache, Projekt, Geschichte*. Erst danach kommen die ersten Vollverben: *sagen* (in der Form *sagte*), *fragen* (*fragt*), *stehen* (oft in metaphorischer Verwendung).
- In einem Fachtext haben naturgemäß (gleich nach den Funktionswörtern) die Bezeichnungen der „Objekte", über die im Text geschrieben wird, die höchste Frequenz.

Mit solchen simplen Recherchen können die Lerner/-innen selbst erfahren, wie die Verteilung einzelner sprachlicher Elemente in einem beliebigen Text aussieht. Es empfiehlt sich am Anfang einen kürzeren und vertrauten Text als Versuchskorpus zu wählen. Im Unterricht eignet sich dieses Instrument 1) für die statistische linguistische Arbeit: Häufigkeiten, typisches Umfeld der KWIC (einfache Sortierung der Konkordanzen nach dem linken oder rechten Kontext); 2) für das Erfassen der „Hauptsäulen" – wichtigsten Träger der Informationen im Text; 3) für das Erschließen von Wortfamilien im gespeicherten Text/ in gespeicherten Texten. Darüber hinaus kann auch gezeigt werden, welche Wörter (Wortarten) in jedem Text gleich sind, welche nicht und warum.

5.2 Linguee

Linguee ist eine „Kombination aus einem Wörterbuch und einer Suchmaschine", wie es die Autoren auf der Homepage (**www.lingue.de → Über Linguee**) nennen. Die Textsammlung besteht aus mehreren Millionen zwei- und mehrsprachigen (fast ausschließlich Internet)Texten. Bisher sind Parallelen Deutsch-Englisch, Deutsch-Französisch, Deutsch-Spanisch und Deutsch-Portugiesisch zugänglich und auch Versionen der Sprachen untereinander (z.B. Spanisch – Englisch, Spanisch – Französisch, Spanisch – Portugiesisch). Die Autoren planen auch Erweiterungen mit anderen Sprachen: Chinesisch, Japanisch und Russisch.

Zugang und Funktionen:
Der Zugang ist kostenlos über www.linguee.de.
Einstellung der Sprachen: mit dem Pfeil neben Default Deutsch⇔Englisch

Suchmöglichkeiten im Linguee:

Wortform: **Suchfeldeingabe: "xxxxx"** → Suchen
Wortkombination: Suche nach genauen Wortformen **Suchfeldeingabe: "xxxxx Yyyyy"** → Suchen Suche nach allen Formen **Suchfeldeingabe: xxxxx Yyyyyy** → Suchen
Lemma: **Suchfeldeingabe: xxxxx** → Suchen

Eingabe	Ergebnis
"blaueres"	*blaueres* (nur diese Form – siehe Abb. 81)
"blauer Himmel"	*blauer Himmel* (nur diese Formen – siehe Abb. 82)
blau	*Blau, blaue, blauen* (alle Formen – siehe Abb. 83)
blau Himmel	*blaue Himmel, blauen Himmel, Blau des Himmels* (alle Formen beider Wörter – siehe Abb. 83)

Die Suche nach Wortteilen ist in Linguee nicht möglich.

Beispiele:

Suche nach der genauen Wortform *blaueres*

- **Suchfeldeingabe: "blaueres"** → Suchen

Die gesuchte genaue Wortform muss (wie bei üblichen Suchmaschinen) in Anführungszeichen gesetzt werden.

Eine "niedrige" Farbtemperatur wie 3200 °C impliziert zum Beispiel ein wärmeres (gelberes/röteres) Licht, während eine "hohe" Farbtemperatur wie 9300°K ein kälteres (**blaueres**) Licht impliziert.	For example, "low" color temperature, like 3200°K, implies warmer (more yellow/red) light, while "high" color temperature, like 9300°K, impliesa cooler (more blue) light.

Abb. 81: Konkordanz zur Abfrage: Wortform *blaueres* (Linguee 2014)

Suche nach der genauen Wortkombination *blauer Himmel*

- **Suchfeldeingabe: "blauer Himmel"** → Suchen

Die Maschine zeigt nur die genaue Kombination *blauer Himmel* an, wie aus der Abb. 82 auf der folgenden Seite ersichtlich ist.

Im Hintergrund erscheint ein Streifen **blauer Himmel** mit oben links der Sonne, die auf die Feigen scheint.	A strip of blue sky appears in the background with the sun and its rays on the left at the top, pointing to the figs.
Blauer Himmel, weißer Sandstrand, Touristen im Liegestuhl mit einem kleinen technischen Wunderkasten vor sich.	Blue skies, white sandy beaches and tourists in deck chairs with a small magical gadget in their laps.

Abb. 82: Konkordanzen zur Abfrage: genaue Phrase *blauer Himmel* (Linguee 2014)

Suche nach einer Wortkombination mit verschiedenen Formen

- **Suchfeldeingabe: blau Himmel** → Suchen

Die Maschine gibt die Kombination *blau* und *Himmel* in unterschiedlichen Flexions- und sogar Ableitungsformen wieder. Die Texte sind also teilweise und sehr einfach lemmatisiert. Ins Suchfeld muss die Grundform eingegeben werden.

Das mittlere Gelbfilter 022 und das dunkle 023 schwächen Blau jeweils noch etwas mehr; der **blaue Himmel** wird entsprechend stärker abgedunkelt.	The medium yellow filter 022 and the dark 023 attenuate blue a little more in each case; the bluesky is made correspondingly darker.
[…] Großstädte wie Salzburg, Wien und Linz und in der Schweiz das steil in den klaren, **blauen Himmel** ragende Matterhorn sowie malerische Dörfer und Seestädte der Weltklasse besuchen.	[…] great cities: Salzburg, Vienna and Linz, combined with Switzerland's steep Matterhorn jutting into the crisp blue sky, its quaint villages and world-class lakeside cities.
…] Weizenfelder sich mit wildem Mohn und Lilien schmückten und der See das unendliche **Blau des Himmels** reflektierte; im Herbst, als der erste Rauche aus den Kaminen stieg, die […]	[…] were interspersed with wild poppies and lilies and the lake reflected the blue of a cloudless sky;in the fall, when the first smoke curled out of the chimneys, the […]

Abb. 83: Konkordanzer zur Abfrage: Lemmata *blau* und *Himmel* (Linguee 2014)

Dieses Instrument eignet sich tatsächlich sehr gut als eine Wörterbuchhilfe und ist zuverlässiger als übliche Übersetzungsprogramme im Internet (etwa google translator). Es ist schnell und – glaubt man der Köllner Mannschaft von Linguee – auch im Angebot der lexikalischen Äquivalente (von Linguee selbst „Redaktionelles Wörterbuch“ genannt) relativ zuverlässig.

Es handelt sich jedoch um kein Korpus (es wird auch nicht als solches deklariert), denn die Texte lassen sich nach keinen Kriterien abgrenzen oder auswählen. Es lässt sich auch nicht feststellen, was die Originalsprache des Textes war, wer die Autoren/-innen waren, aus welchem Bereich der Text stammt etc. Bei jeder Konkordanz ist jedoch ein Link zur Homepage der Texte. Andere Annotationen (Metadaten, Tagging) sind nicht vorhanden.

5.3 Österreichisches Aussprachewörterbuch, Österreichische Aussprachedatenbank (ADABA)

Das **Österreichische Aussprachewörterbuch** und die **Österreichische Aussprachedatenbank** (ADABA) sind an der Universität Graz unter der Leitung von Rudolf Muhr entstanden und stellen (wie aus dem vollen Namen ersichtlich ist) eine Kombination von einem Wörterbuch und einer Datenbank das. Die ADABA hat viele Eigenschaften üblicher Korpora, es ist aber kein Korpus: Es beinhaltet keine natürlichen Texte, weiterhin

werden die Ergebnisse nicht in Form von Konkordanzen präsentiert und es sind auch keine statistischen Funktionen vorhanden. Das Ziel dieses Instruments wird deutlich auf der Homepage bestimmt:

> „Mit dem Österreichischen Aussprachewörterbuch (ÖAWB) und der damit verbundenen Österreichischen Aussprachedatenbank (ADABA) steht erstmals eine umfassende Dokumentation der Aussprache des Österreichischen Deutsch zur Verfügung. ADABA und ÖAWB dokumentieren die verschiedenen Aussprachеformen und Standardvarianten in Österreich und bieten umfassende Hilfen für die Ausspracheschulung an. Die hier dargestellte Aussprache beschreibt die derzeit in Österreich übliche «Medienpräsentationsnorm», die jener Deutschlands und der Schweiz gegenübergestellt wird. Wörter und Texte liegen transkribiert vor und können beliebig oft nach vielen verschiedenen Kriterien abgehört werden. Umfassende Analysen der Unterschiede sowie eine genaue Beschreibung der Funktionalität der ADABA ergänzen das ÖAWB."
> (adaba.at: Aussprachewörterbuch → 1. Daten im Überblick → Ziele. 1.2.2014)

Zugang und Funktionen

Der Zugang ist kostenlos über **http://adaba.at/.**

ADABA-WEB Auswahl →

1. Wörterbuch
2. Texte (Textdatenbank)
3. Hilfe (übersichtliches kurzes Manual

1. Wörterbuch

Suchmöglichkeiten im Aussprachewörterbuch:

Orthographische Suche	Suche nach der üblichen Orthographie
Phonetische Suche	Suche über phonetische Transkriptionscodes. Bei dieser Auswahl öffnet sich das Fenster mit der SAMPA – IPA-Tabelle.
ganzes **Wort:**	**Suchfeldeingabe:** (gewünschtes Wort) → 🔎
Wortteil:	
Suche nach dem Wortanfang:	☑ anlautend auswählen
Suche nach dem Wortende:	☑ inlautend auswählen
Suche nach dem Wortende	☑ auslautend auswählen
Suchfeldeingabe: (Anfang, Mitte oder Ende eines Wortes) → 🔎	

Beispiele:

Suche nach der Aussprache von *Mathematik*

Orthographische Suche		auswählen
Suchoptionen: Wörterbuch-Auswahl → Audiokorpus		auswählen
Suchfeldeingabe: Mathematik	🔎	anklicken
Das gesuchte Wort (Mathematik) im Kästchen neben den Flaggen		anklicken
Blauen/ Roten Pfeil neben der gewünschten Varietät zum Abhören		anklicken

Suche nach der Aussprache von *Ch-* am Anfang der entlehnten Substantive

Orthographische Suche		auswählen
Suchoptionen: Wörterbuch-Auswahl → Audiokorpus		auswählen
Wortkategorie:		auswählen
Grammatisch →	Substantiv	
Etymologisch →	Lehnwort	auswählen
Lautumgebung →	anlautend	auswählen
Suchfeldeingabe: ch		anklicken

Die Ergebnisse der Suche (*Cha-Cha-Cha*, *Chaiselongue* bis *Chronik*) erscheinen im Kästchen neben den Flaggen. Nach dem Anklicken des Wortes kann man die Aussprache anhören (über den blauen/ roten Pfeil neben der jeweiligen Varietät).

2. Texte (Textdatenbank)

ADABA beinhaltet auch kohärente (gesprochene) Texte mit Transkripten. In den Texten lässt sich nach Texttyp (biographischer Text, Nachrichtentext etc.) und Sprecher/in (Deutschland männlich/ weiblich, Österreich männlich/ weiblich; Schweiz männlich/ weiblich) recherchieren. Zum gesuchten Text erscheint die genormte und die phonetische (IPA) Transkription. Eine Erstellung von Konkordanzzeilen ist nicht möglich.
Dieses Instrument ist hilfreich für die „plurizentrische" Phonetik bzw. auch als Trainingsportal für die Aussprache in einzelnen deutschsprachigen Ländern.
Da die Suche entweder orthographisch oder phonetisch erfolgen kann (phonetisch sind die Texte in IPA[25]-Zeichen transkribiert), eignet sich ADABA auch zu Diskussionen und Übungen in Phonologie.

Es empfiehlt sich ADABA in unterschiedlichen Web-Browsern auszuprobieren (empfohlen: Mozilla Firefox), da einige Browser (Google Chrome) die Transkripte schlecht oder gar nicht wiedergeben können.

5.4 ParZu

Ebenfalls kein Korpus, sondern eine Software und ein hilfreiches Instrument zum Demonstrieren, wie die Analyse der Sprache automatisch funktioniert, ist der Parser ParZu, der am Institut für Computerlinguistik der Universität Zürich entwickelt wurde (Sennrich 2009).

Zugang:
Auf der Adresse http://kitt.cl.uzh.ch/kitt/parzu/ erscheint das **ParZu - The Zurich Dependency Parser for German.**

Ins Abfragefenster Input text to parse: kann man einen beliebigen Text schreiben oder einfügen. Empfehlenswert (wegen der Übersichtlichkeit) sind maximal zwei einfache Sätze oder ein komplexer Satz.

[25] Internationales phonetisches Alphabet

Beispiel:
Mit dem Parser wird ein Satz aus der österreichischen Tageszeitung *Kurier* analysiert:
Stadt Wien weist den Vorwurf der illegalen Parteienfinanzierung zurück (Kurier 2014).

Den Satz: ***Stadt Wien weist den Vorwurf der illegalen Parteienfinanzierung zurück.***	ins Abfragefenster kopieren
Output format: Graphical (first sentence only	auswählen
	SEND anklicken

Das Ergebnis ist eine Graphik (ein Dependenzbaum), aus der die innere Satzstruktur ersichtlich ist – die zugewiesenen Rollen erscheinen als Beschriftungen der Linien. In der Abb. 84 auf der nächsten Seite sind es: subj., objа, avz, app, det, gmod, attr.

Der verbale Teil des Prädikats (*weist*) ist das Zentrum des Satzes und wird nicht beschriftet.

Die Beschreibung der Abb. 84 von links nach rechts:

subj: Subjekt *Stadt* mit Apposition (**app**.) *Wien*
objа: Akkusativobjekt *(den) Vorwurf* modifiziert durch eine Genitivergänzung (**gmod**) *(der) illegalen Parteienfinanzierung* (genitive modifier (Sennrich et al 2009: 21) mit dem Attribut (attr) *illegalen*)
avz: abgetrennter Verbzusatz *zurück.*

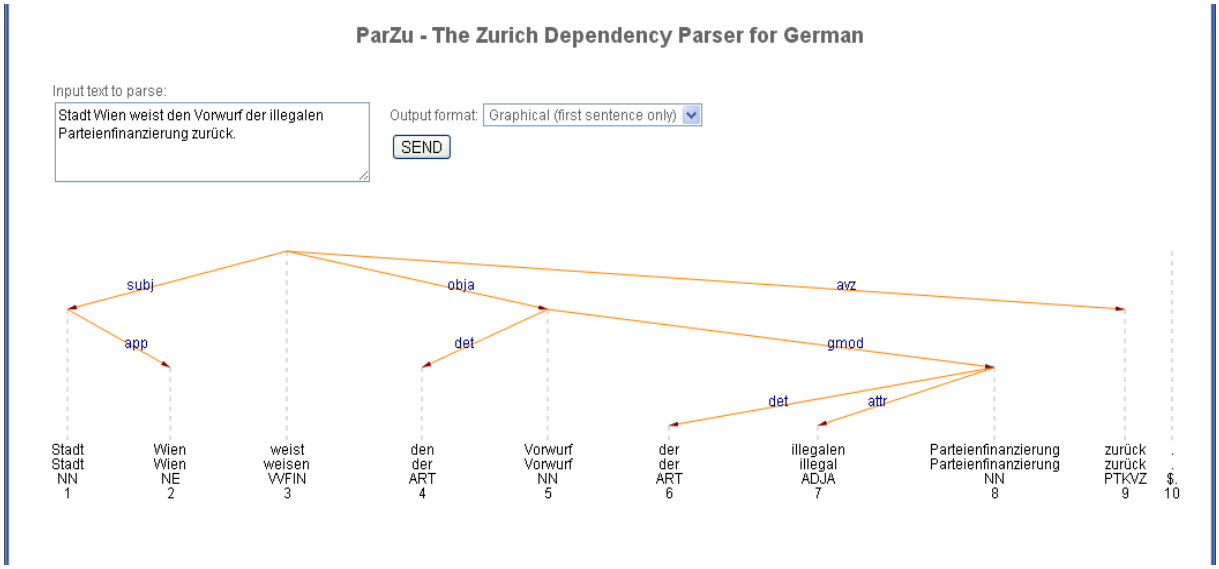

Abb. 84: Geparster Satz: *Stadt Wien weist den Vorwurf der illegalen Parteienfinanzierung zurück*. (ParZu)

Darüber hinaus sind in der Graphik auch die Lemmatisierung und das Tagging beispielhaft zu sehen:

Form	Stadt	Wien	weist	den	Vorwurf	der	illegalen	Parteifinanzierung	zurück
Lemma	Stadt	Wien	weisen	der	Vorwurf	der	illegal	Parteifinanzierung	zurück
Tag	NN	NE	VVFIN	ART	NN	ART	ADJA	NN	PTKVZ

Erklärung der Tags:

NN	(normales) Nomen
NE	Nomen – Eigenname
VVFIN	Vollverb, finite Form
ART	Artikel
ADJA	Adjektiv in attributiver Position
PTKVZ	abgetrennter Verbzusatz

Die Tag-Codes sind ident mit dem Stuttgarter Tagset (siehe Kap. 8.2).

ParZu eignet sich gut als Einstieg zur Erklärung einiger Korpustools und Prozesse: Lemmatisierung, Tagging, Parsing.

Mehr zu geparsten Korpora und zu den zugänglichen Treebanks bringen Jelínek et al. (2012) und Rosen (2012).

6. Korpusarbeit – Studien

Der Einsatz von Korpora und anderen elektronischen Instrumenten im Sprachunterricht, besonders aber in DaF/DaZ, hängt mit mehreren Überlegungen zusammen (dazu auch Heine 2008: 4), die in folgenden Fragen zusammengefasst werden können:

1) Ist es überhaupt sinnvoll, Korpora und ähnliche Instrumente im (Fremd-)Sprachenunterricht einzusetzen?
2) Wenn es sinnvoll ist, wem soll das Korpus/ Instrument dienen - den Lernenden oder den Unterrichtenden?
3) Hat man dafür Zeit?
4) Wenn ja, wie viel Zeit ist man bereit der Korpusarbeit zu widmen?
5) Gibt es dazu auch die entsprechende technische Ausrüstung (Computerraum, Internetzugang)?
6) Wie gewandt sind alle Betroffenen (die Lernenden, aber auch die Lehrenden) im Umgang mit der Technik?

Zum ersten Punkt sind schon einige Publikationen und Artikel erschienen. Bereits die ersten davon, u.a. Leech (1997: 22-23), betonten, dass der Korpuseinsatz den kommunikativen Sprachunterricht positiv beeinflussen kann. Er kann ihn durch gut überlegte Aufgaben unterstützen und um einige Dimensionen erweitern. Im Falle des Fremd- und Zweitsprachenunterrichts ist dies beispielsweise die Interkulturalität, die durch parallele Texte automatisch mittransportiert wird. Die Korpusdaten können nämlich Gemeinsamkeiten und Unterschiede im kulturbedingten Sprachgebrauch aufdecken, neue Informationen über die andere Sprache und über das Zielland bringen. Es muss noch einmal betont werden, dass Texte in Korpora, die hier angesprochen werden, ausschließlich aus der realen Welt kommen. Diese Tatsache mache wiederum die Arbeit mit Korpustexten in den Anfangsphasen des Deutschlernens unmöglich, beklagen viele Elementaristen/-innen, da die authentische Sprache oft zu kompliziert sei. Dies mag immer wieder zutreffen, aber wenn man den Lernenden am Anfang die Rolle eines Beobachters/ einer Beobachterin der Sprache zuweist und zu einfachen Aufgaben motiviert (etwa wie wird (wo) etwas ausgesprochen, geschrieben; welche Formen gibt es von einem Grundwort etc.), könnte es ein guter Einstieg in die spätere selbstständige Arbeit mit einem Korpus bedeuten. Darüber hinaus ist bekannt, dass man authentischen Texten weitere, unerwartete und interessante Informationen entnehmen kann.

Aus organisatorischen, zeitlichen aber auch pragmatischen Gründen ist der Korpuseinsatz *für* den Unterricht sichtlich bedeutender als *im* Unterricht selbst, zumindest in den Anfangsphasen. Reflektierte Lehrkräfte sollten zu Korpora greifen. Wie sonst könnten sie einzelne sprachliche Phänomene, Regularitäten oder Ausnahmen den Lernenden erklären und belegen? Das „Bauchgefühl" funktioniert bei muttersprachlichen und/ oder erfahrenen Lehrern/-innen zwar einwandfrei. Besser ist es jedoch, wenn dieses „Gefühl", bzw. längst in Vergessenheit geratene Schulformeln, gleich auch durch mehrere Belege objektiviert werden. Das Sprachgefühl kann eigentlich jede/-r mit ausreichend großen Korpora verifizieren (vgl. Studie 9). So gesehen bedeuten plurizentrisch ausgewogene Korpora auch eine große Chance für den (endlich konsequent umgesetzten) plurizentrischen Ansatz in der Deutschvermittlung.

Der Zeitaufwand muss dabei nicht so groß sein, wenn man sich ein Bild über die Korpuslandschaft und Möglichkeiten der Korpusmanager verschafft hat. Zuerst muss man natürlich mit einem (wenn auch minimalen) Zeitaufwand rechnen. Je besser jemand ein Korpus nutzen will, desto mehr Zeit muss er/sie sich für die Recherchen nehmen.

An Murphys Gesetze darf dabei gar nicht gedacht werden! Dass Korpusmanager auf einmal streiken, Korpus-Server, PC oder Internet abstürzen - immer wieder Pannen auftreten, kann niemand vermeiden. (Nur gegen Stromausfall kann man sich durch Verwendung eines Notebooks rüsten.) Nichtsdestotrotz ist eine Recherche im Korpus des Aufwands wert, wenn man bedenkt, wie lange eine Suche nach einem bestimmten Wort in mehreren Büchern dauern würde.

Die subjektiv-technischen Punkte müssen ad hoc entschieden werden. Zu bemerken ist, dass es (besonders bei jüngeren Schülerinnen und Schülern) Fälle gibt, in denen ein in den Sprachen eher schlechter Lernende das Interesse an einer Sprache über EDV-Instrumente gewonnen hat. Anfängern wird jedoch empfohlen, sich zuerst ein einfacheres Instrument auszusuchen (**Wortschatz**, **ADABA** oder **DWDS**), einige Abfragen auszuprobieren und so in den Möglichkeiten der Korpusrecherche zu schnuppern.

In den folgenden Studien werden einige Ideen präsentiert, wie man Korpora zur Lösung ausgewählter sprachlicher Fragestellungen nutzen kann. Sie gehen von den erfahrungsgemäß häufigsten Problemen aus, die Lerner/-innen mit unterschiedlichen (europäischen) Erstsprachen beim Erlernen der deutschen Sprache haben. Die Problempunkte verfolgen ungefähr die „klassische" strukturalistische Linie (in der ich eine gewisse Logik beim ganzheitlichen Er-/ und Begreifen einer Sprache sehe), ausgehend von der kleinsten Einheit. Dies ist übrigens keine originelle Idee. Man findet sie bei Lemnitzer/ Zinsmeister (2010: 124-167) für Deutsch, genauso aber auch bei Čermák et al. (2005) für Tschechisch. Während Lemnitzer/ Zinsmeister die theoretische Grundlage für die linguistischen Recherchen präsentieren, in denen dem Leser freie Hand überlassen wird, in welchem Korpus recherchiert werden kann, bieten die Autoren des zweiten Buchs konkrete Beispiele an, wie die Syntax der Abfrage aussieht, wie man Konkordanzen sortiert usw. Das Buch ist als ein Übungsbuch (mit Schlüssel) konzipiert und wird ausschließlich anhand von Daten des Tschechischen Nationalkorpus aufgebaut. Beide Bücher richten sich eher an Linguisten/-innen.

Die Fallstudien hier werden dem DaF/DaZ-Bereich angepasst, betrachten die deutsche Sprache in manchen Punkten plurizentrisch und kontrastiv. Der Vorgang bei den Recherchen wird Schritt für Schritt beschrieben. Jede Studie bringt Ergebnisse, die bisher gar nicht oder zumindest statistisch noch nicht erfasst worden sind. Im Anhang werden weitere statistische Angaben über die deutsche Sprache angeführt, die keinen Kommentar benötigen. Jede der folgenden 13 Studien zeigt einen oder mehrere Lösungswege zu Fragestellungen auf mehreren sprachlichen Ebenen auf. Sie gehen dabei nicht ins letzte Detail, weil sie als Beispiele für die Korpusarbeit und als Inspiration für weitere oder ähnliche Untersuchungen dienen sollen, die in der Erstellung von Lehrmaterialien und Impulsen für den Unterricht münden.

Die Aufgaben und Übungen, welche mithilfe von Korpusdaten erstellt werden können, lassen sich grundsätzlich in diese Typen zusammenfassen:
1) induktive Aufgaben (Feststellen von Regeln)
2) deduktive Aufgaben (Suchen nach Belegen für gewisse Regel)
3) Einsetzübungen
4) Zuordnungsübungen (z.B. Zuordnung passender Überschriften zu Textpassagen)

Die inhaltliche Vielfalt der Aufgaben und Übungen ist jedoch so breit, wie die Sprache reich und bunt ist.

Phonetik und Orthographie

Das lautliche und das optische Bild der Sprache wird oft mit der Fassade eines Gebäudes oder mit der Bekleidung eines Menschen verglichen. Wie wichtig oder unwichtig diese „Gewänder“ für die Verständigung und demnach auch für den DaF/DaZ-Unterricht sind, wird hier nicht diskutiert. Dass sie im Fremdsprachenerwerb immer eine gewisse Rolle einnehmen werden, bleibt jedoch unumstritten. Einige Korpora und korpusähnliche Instrumente können helfen, diese Ebene zu beobachten, auf Unterschiede hinzuweisen oder Lücken in Lehrbüchern zu schließen.

Studie 1: Suche nach der „richtigen“ Aussprache

Eine richtige Aussprache gibt es in keiner Sprache. Die des Deutschen hat dazu drei Varianten, alle drei gelten als Norm. Obwohl die Erstellung elektronischer Instrumente für die Aussprache technisch aufwendig ist, gibt es bereits auch auf diesem Gebiet einige Werkzeuge, die ein einigermaßen plastisches, wenn auch immer noch nicht repräsentatives Bild über die Aussprache des Deutschen darstellen können.

Die „geschulte“ Aussprache (einzelner Wörter) des Deutschen ist aus mehreren elektronischen Quellen abrufbar. Eine Tonspur hat auch fast jedes elektronische Wörterbuch. Die Qualität der Wörterbücher (allgemein) lässt oft zu wünschen übrig, dies gilt auch für die Aussprache. Die Objektivität erreicht man durch Vergleich von mehreren Quellen, deswegen ist es ratsam auch für die Phonetik mehrere Instrumente auszuprobieren.
Die (rein bundesdeutsche) Aussprache findet man im **DWDS**, für den plurizentrischen Vergleich empfiehlt sich **ADABA** und Beispiele für die nicht-geschulte, spontane Aussprache in ganzen Texten findet man in der **DGD**.

Fragestellung 1: *Wie ist die Betonung im Wort* Kaffee, Sakko, Mathematik*? Wie wird im Auslaut* -ig *ausgesprochen? Gibt es regionale Unterschiede in der Aussprache der Fremdwörter, die auf* -ie *enden?*

Recherchiert wird im **DWDS**, **ADABA** und in der **DGD**. Die Ergebnisse werden dann verglichen.

Recherche im DWDS

Zuerst wird die **Betonung der Wörter** *Kaffee, Sakko* und *Mathematik* untersucht.

- **DWDS** → Abfragefenster
- Suchfeldeingabe: **Kaffee** → Suche im DWDS
  dann: **Sakko** →
  dann: **Mathematik** →
- **DWDS-Wörterbuch:**
- **Aussprache:** ➢ (Pfeil anklicken)

Die Tonspur ist nicht zu allen Lexemen vorhanden. Die Sprecher sind ausschließlich aus dem deutschsprachigen Raum (allesamt ARD-Mitarbeiter/-innen).

Obwohl bei einigen wenigen Lexemen mehrere Tonspuren abrufbar sind, handelt es sich stets um eine bundesdeutsche Färbung der Aussprache. Dies erkennt man auch beim Anhören der Wörter *Kaffee* und *Sakko*, die jeweils zwei Tonspuren haben:

Kaffee: [deɐ ˊkafe:], [deɐ kaˊfe:]

Der Unterschied im Wortakzent (*Kaffe – Kaffe*) wird nicht erklärt.

Sakko: [deɐ ˊzako:], [das ˊzako:]

Der Wortakzent (nur „bundesdeutsch" auf der ersten Silbe: *Sakko*) bleibt gleich, angedeutet wird das schwankende Genus.

Mathematik: [matemaˊti:k^{h}]

Hier gibt es nur eine Tonspur mit der bundesdeutschen Standardbetonung auf der letzten Silbe (*Mathematik*).

Das ***-ig* im Auslaut** wurde probeweise an drei Wörtern im **DWDS** überprüft: *schwierig*, *lustig* und *zwanzig*.

- **DWDS** → Abfragefenster
- **Suchfeldeingabe**: **schwierig** → 🔍
  dann: **lustig** → 🔍
  dann: **zwanzig** → 🔍
- **DWDS-Wörterbuch:**
- **Aussprache:** ➢ (Pfeil anklicken)

Alle *-ig* im Auslaut werden nur als „ich-Laut" ausgesprochen:

schwierig: [ˊʃvi:ʀiç]
lustig: [ˊlustiç]
zwanzig: [ˊtsvantsiç]

Von einer plurizentrischen Abdeckung kann im DWDS kaum gesprochen werden. Dies könnte sogar Absicht sein, denn ein Teil des DWDS wird zum deutschen Teil des plurizentrischen Korpus C4 (siehe dazu Kap. 4.2.1.3).

Die Suche nach den Unterschieden in der **Aussprache der Fremdwörter**, die **auf *-ie*** enden, wäre im DWDS nicht sinnvoll. Im Korpus kann man zwar alle auf *-ie* endenden Wörter abrufen (Suchfeldeingabe: *ie), die Auswahl der Kandidaten für die weitere Recherche wäre jedoch sehr umständlich, denn das Korpus liefert über 2,7 Mio. Belege (u.a. natürlich auch die Wortformen *die*, *sie*, *wie*, die im Deutschen zu den am meisten verwendeten Wörtern gehören). Für diese Aufgabe ist es deswegen sinnvoll, ein anderes Instrument zu verwenden.

Ein plurizentrisches Instrument für die Phonetik und Phonologie ist die **ADABA** (Österreichische Aussprache Datenbank/ Österreichisches Aussprachewörterbuch). Alle Sprecher/-innen sind „geschulte" professionelle Radiomoderatoren/-innen der öffentlich-rechtlichen Anstalten aller drei deutschsprachigen Länder (ARD, ORF und SRG) und decken somit alle Varietäten des Deutschen ab. Das abgerufene Wort ist in jeder Varietät auch mit der IPA-Transkription versehen. Den Lernenden kann vor allem das Aussprachewörterbuch behilflich sein.

Recherche in der ADABA

Zuerst wird die **Betonung der Wörter** *Kaffee, Sakko* und *Mathematik* nacheinander abgefragt:

- **ADABA** → ADABA-WEB-Auswahl → Wörterbuch → Orthographische Suche
- **Suchfeldeingabe**: **Kaffee** → 🔍
  dann: **Sakko** → 🔍
  dann: **Mathematik** → 🔍

Das Wort erscheint im Kästchen neben den Landesflaggen und muss angeklickt werden. Erst dadurch wird die Transkription und Tonspur abgerufen.

Die Aussprachedifferenzen sind in die Tab. 20 zusammengefasst:

	Kaffee	*Sakko*	*Mathematik*
D:	[ˊkafe:]	[ˊzako:]	[matemaˊti:k^{h}]
A:	[kaˊfe:]	[saˊko:]	[mateˊma:tik]
Ch:	[kaˊfe:]	[ˊsako]/[ˊza..]	[matemaˊti:k^{h}]

Tab. 20: Regionale Unterschiede in der Aussprache *Kaffee, Sakko, Mathematik*

Das Instrument hat auch einige „Korpuseigenschaften[26]“ – es kann z.B. nach Wortteilen gesucht werden (siehe Kap. 5.3). Deswegen kann leicht auch die **Aussprache von *-ig* im Auslaut** abgefragt werden (unter Such-Optionen auslautend auswählen. Dazu auf der folgenden Seite die Beschreibung von *-ie* im Auslaut). Hier werden drei volle Wörter verglichen:

- **ADABA** → ADABA-WEB-Auswahl → Wörterbuch → Orthographische Suche
- Wörterbuch-Auswahl: Audiokorpus auswählen
- **Suchfeldeingabe:** **schwierig** → 🔍
  dann: **lustig** → 🔍
  dann: **zwanzig** → 🔍

Die Ergebnisse sind klar zu sehen und (nach anklicken der Audiofiles) auch zu hören. Zusammengefasst sind sie in der Tab. 21:

	schwierig	*lustig*	*zwanzig*
D:	[ˊʃvi:Riç]	[ˊlustiç]	[ˊtsvantsiç]
A:	[ˊʃvi:Rik]-[..Riç]	[ˊlustik]-[..iç]	[ˊtsvantsik]-[..siç]
Ch:	[ˊʃvi:Rik]-[..Riç]	[ˊlustik]-[..iç]	[ˊtsvantsik]-[..siç]

Tab. 21: Regionale Unterschiede in der Aussprache *-ig* im Auslaut: *schwierig, lustig, zwanzig*

Eine interessante Erweiterung dieser Aufgabe (zum Selbststudium) bietet auch die Sortierung der Wörter, die auf *-ig* enden (unter Such-Optionen auslautend auswählen, ins Suchfeld wird **ig** eingegeben). Unter ihnen befinden sich eben auch *Bahnsteig, Braunschweig, Bürgersteig, Gehsteig, Steig, Teig* und *Zweig*. Die Lernenden können selbst den Unterschied in der Aussprache zwischen dem Monophthong [i] und dem Diphthong [a͡e] gefolgt von *g* erkunden.

[26] Eigentlich ist **ADABA** kein Korpus, weil nicht in Texten gesucht wird, es sind keine Konkordanzen abrufbar, es gibt auch keine statistischen Funktionen.

Auch für die Erkundung der **Aussprache von *-ie* im Auslaut** müssen nur die Such-Optionen eingestellt werden (keine Platzhalter sind notwendig):

- **ADABA** → ADABA-WEB-Auswahl → Wörterbuch → Orthographische Suche
- Such-Optionen**:**

  Grammatisch: Substantiv auswählen
  Lautumgebung: auslautend auswählen
  Wörterbuch-Auswahl: Audiokorpus auswählen
- **Suchfeldeingabe: ie** →

Das Ergebnis der Abfrage sind etwa 130 Substantive (von *Aborigine* bis *Xenie*), die man jetzt manuell durchgehen muss. Der Unterschied in der Aussprache ist gleich beim ersten Wort deutlich zu sehen:
Aborigine:

D: [ɛboˊʁi:dʃi:ni:]/[..ʧi:ni:]
A: [ɛboˊʁiʧi:ni:]/[..ˊʁidʤi:ni:]
Ch: [aboˊri:ʧi:ne:]/[…ʤi:ne:]

Ein markanter Unterschied ist auch im Wort *Serie* zu erkennen:

D: [ˊze:ʀi.e]/[..ʀji.e]
A: [ˊse:ʀije:]
Ch: [seˊʀi:]

In der Schweiz wird *Serie* mit auslautendem […i:] ausgesprochen, im bundesdeutschen und österreichischen Standard mit […ije].
Es ist zu bemerken, dass es sich dabei um Standardaussprache handelt. In der schweizerischen Umgangssprache werden z.B. auch das Wort *Studie* und gelegentlich auch *Familie* mit […i:] ausgesprochen.

Recherche in der DGD

Die Rechercheergebnisse aus Instrumenten, die die Standardaussprache repräsentieren, kann man auch mit Aufnahmen der (z.T. spontan) gesprochenen Sprache in der DGD überprüfen. Zuerst werden wieder die Wörter *Kaffee*, *Sakko* und *Mathematik* abgefragt, die Audiofiles abgehört und die Betonung notiert:

- **DGD** → Recherche → Tokens
- Korpuswahl: alle einschalten
- **Suchfeldeingabe:** Lemma: **Kaffee** → Suche starten
  dann: Lemma: **Sakko** → Suche starten
  dann: Lemma: **Mathematik** → Suche starten

Das Wort *Sakko* findet man in der DGD (momentan) nicht.

Die meisten Sprechereignisse (gesprochene Texte) bilden Aufnahmen von Sprecher/-innen aus Deutschland. Die Aussprache repräsentiert den bundesdeutschen Standard [matemaˊti:k^{h}] bis Substandard [ˊkafe][27].

[27] Standardisiert: [ˊkafe:], [kaˊfe:]

Des Weiteren wurde die **Aussprache von *-ig* im Auslaut** abgefragt:

- **DGD** → Recherche → Tokens
- Korpuswahl: alle einschalten
- **Suchfeldeingabe:** Normalisiert: **%ig** → Suche starten

Die Abfrage nach *-ig* am Ende eines Wortes liefert interessante Ergebnisse: An manchen Stellen ist die Schreibweise nicht „normalisiert", sondern nur transkribiert (siehe Abb. 85).

FOLK_S_00030	wie	**eklich**	sagt oma
FOLK_S_00026	skop dran voltmeter dran durschtreten guck isch	**fertisch**	aus die maus
FOLK_S_00026	net isch isch bin do leischt	**linkslastisch**	ja
FOLK_S_00026	glaub isch null komma fünf zwo gell	**rischtisch**	
FOLK_S_00026	wenn s n	**richtich**	gutes steurgerät wär en grössere teil
FOLK_S_00026	okay awwer	**rischtich**	was der herr fischer gsaht hot
FOLK_S_00026	mir war	**wischtisch**	dass mar den hallgeber prüfen

Abb. 85: Konkordanzen zur Abfrage **%ig** (DGD)

Auch aus diesen Belegen ist ersichtlich, dass die Sprecher/-innen im größten Teil des Korpus eindeutig das bundesdeutsche Idiom sprechen. Auch bei den meisten Belegen, die mit auslautendem *-ig* transkribiert wurden (*richtig*, *eklig*, *fertig*), ist die Aussprache von *-ig* eindeutig als ein Ich-Laut zu hören.

Zum Schluss wurde noch die **Aussprache von *-ie* im Auslaut** abgefragt:

- **DGD** → Recherche → Tokens
- Korpuswahl: alle einschalten
- **Suchfeldeingabe:** Normalisiert: **%ie** → Suche starten

Eine Abfrage nach allen Wörtern, die auf *-ie* enden, erwies sich als wenig sinnvoll. Das Korpus liefert nämlich über 38 tausend Belege, darunter auch *die, sie, nie* und *(irgend)wie*. Das Filtern ist recht mühsam, daher ist es einfacher, direkt die Wörter, die im **ADABA** gefunden wurden, zu suchen: *Aborigine* ist nicht in der DGD vertreten. Das Wort *Serie* kann man in 25 vertonten Treffern anhören, ausschließlich von Sprechern und Sprecherinnen aus Deutschland, die Aussprache lautet [ˊze:ʀi.e], bzw. auch [ˊse:ʀi.e].

Fazit:
Mithilfe elektronischer Instrumente lassen sich Unterschiede in der Aussprache einzelner Wörter auf der segmentalen (hier *-ig, -ie*) und auch auf suprasegmentalen (hier Wortakzent) Ebene gut demonstrieren. Diese Unterschiede ließen sich auch quantitativ beweisen, wenn es ein regional ausgewogenes Korpus der gesprochenen Sprache mit Transkripten und/oder Tonspuren gäbe.

Fragestellung 2: *Kann man mit elektronischen Sprachkorpora auch Suprasegmentalia (z.B. Satzakzent, Satzmelodie) untersuchen?*

Vorgang: Wenn Beispiele für die Satzmelodie oder Intonation gesucht werden und die Hörtexte der Lehrwerke nicht ausreichen, bzw. nicht ausreichend ausgewogen sind, kann man sich zweier Instrumente gut bedienen: ganze gesprochene Texte mit abrufbaren Tonspuren sind in der **DGD** und auch in der **ADABA** zu finden.

Recherche in der DGD

Beispiele der Satzmelodie kann man mit beliebiger Abfrage abrufen und die Tonspur anhören. Als gute Beispiele zeigten sich diese Abfragen: 1) Wort *danke*; 2) Verbindung *Schule* und *gehen*.
Abfrage nach dem Wort *danke*:

- **DGD** → Recherche → Tokens
- Korpuswahl: FOLK
- **Suchfeldeingabe:** Normalisiert: **danke** → Suche starten

Die Ergebnisse erscheinen in Konkordanzzeilen, mit dem KWIC in der Spalte Treffer.

Ereignis	Audio	Transkript		Treffer	
FOLK_S_00026	➢		gut	**dange**	
FOLK_S_00026	➢		gut	**danke**	
FOLK_S_00027	➢			**danke**	schön
FOLK_S_00027	➢			**danke**	schön
FOLK_S_00027	➢			**danke**	schön sabine
FOLK_S_00027	➢			**danke**	schön
FOLK_S_00027	➢		so	**danke**	schön nina

Abb. 86: Konkordanzen zur Abfrage danke (DGD, FOLK)

Zu jedem Beleg kann man durch Klicken auf den Pfeil ➢ das Audiofile aktivieren und Passagen von je 15 Sekunden anhören. Eigentlich handelt es sich in diesen Beispielen um ganze Sätze.

Abfrage nach der Verbindung *Schule* und *gehen*:
Diese Abfrage nach einer Verbindung bzw. nach mehreren Wörtern muss in der DGD in der Volltextsuche getätigt werden:

- **DGD** → Recherche → Volltext
- Korpuswahl: FOLK
- **Suchfeldeingabe:** **NEAR(($gehen,Schule),4,false)** → Suche starten

Die Abfrage bedeutet: Suche nach dem Lemma *gehen* und der Wortform *Schule* im Abstand von max. 4 Positionen voneinander, Reihenfolge beliebig.

Auch hier erscheinen die Ergebnisse in Form von Konkordanzzeilen (aussortiert aus 263 Treffern):

#	Transkript-ID	KWIC-Liste	Score	Hörprobe
1	FOLK_E_0015(	ja sie **geht** in **schule** ja **...**welche **schule geht** se	56	➢
2	FOLK_E_0018:	jetzt hier in die **schule gehenden** mitschülern zu den	42	➢
3	FOLK_E_0016:	de er damals zur **schule gegangen ...**zur **schule ging** also	29	➢
4	FOLK_E_0002(	genau also es **geht** um **schule** und dann sagt er**...**sie **geht** in die	28	➢
5	FOLK_E_0002-	wie des in der **schule** ab **geht** weil sie hat**...**die **schule geht** und	28	➢
6	FOLK_E_0000-	zehn jahre in die **schule geh** oder immer up	14	➢
7	FOLK_E_0017!	also wenn die **schule** zeitig anfängt **geht** dann nach n	14	➢
8	FOLK_E_0018:	irgendeinem hier in der **schule geht**	14	➢
9	FOLK_E_0002:	er s pät v on der **schule** kommt **geht** s halt net	14	➢
10	FOLK_E_0018!	leben di e **schule gehen** des find ich	14	➢

Abb. 87: Konkordanzen zur Abfrage: Lemma *gehen* und *Schule*, Abstand 4 (DGD)

Die Tonspur wird durch einen Klick auf den Pfeil ▶ in der Spalte Hörprobe abgerufen.

Die Ergebnisse beider Recherchen in der DGD zeigen, dass man sehr schnell zu Ausschnitten vieler gesprochener Texte gelangen kann. Sollten die Konkordanzen mit der Vertonung im Unterricht eingesetzt werden, müssten sie aber sorgfältig aussortiert werden. Viele Tonspuren sind schlecht zu hören, einige falsch mit dem Text aligniert. Und es gibt in der ganzen DGD nur äußerst wenige Texte von Österreichern und Schweizern. Als ein Instrument für die plurizentrische Untersuchung der Satzmelodie ist die DGD nur sehr eingeschränkt geeignet (vgl. Bestand der DGD2 auf ihrer Homepage).

Recherche in der ADABA

Eine Alternative zur DGD bietet **ADABA**, in der einige Texte jeweils in drei Formen gespeichert sind: gesprochen (gelesen) auf der Tonspur, „normal" geschrieben und phonetisch transkribiert. Hier lassen sich die Unterschiede in der Satzmelodie besser demonstrieren.
ADABA:

- **ADABA** → ADABA-WEB-Auswahl → Texte
- Optionen: Texttyp (Auswahl: Biographischer Text bis sonstiges)
  Sprecher/-in (aus unterschiedlichen deutschsprachigen Regionen)

Durch Klicken auf den Pfeil (▶play) beginnt sich das Audiofile abzuspielen.
Im Feld Ergebnis erscheint der entsprechende Text mit seiner phonetischen Transkription:

Ich bin geboren in Wien, hab meine Kindheit und Jugendzeit zunächst vier Jahre in Vorarlberg verbracht und dann in Maria Enzersdorf - das ist ein kleiner Ort südlich von Wien. iç bɪn gɜbo̝ɐn̩ ɪn ˊviːn hab̥ maenəˊ ˊkɪndhaet unt ˊjuːgənt.tsaet tsune̝ːçst fiːɐ jaːʀə ɪn fo̝ɑʀlbeɐk fəˊb̥ʀaxt unt ˊdan ɪn maʀɪa ˊentsɐsdo̝ɐf dɐs ɪst aen klaenɐ o̝ɐt ˊsyːdliç fo̝n ˊviːn

Abb. 88: Bibliographischer Text, Sprecher aus Österreich, männlich (ADABA)

Ich bin geboren in Stuttgart. Dort habe ich aber nur sehr kurz gelebt. iç bɪn gəˊboːʀən ɪn ˊʃtutgaːtʰ dɔɐt haːb iç avɐ nuːɐ seːɐ kuʁts gəˊleːpt

Abb. 89: Bibliographischer Text, Sprecherin aus Deutschland, weiblich (ADABA)

Die Satzmelodie kann man nach Abrufen unterschiedlicher Sprecher/-innen vergleichen. Sie ist jedoch nicht im Text gekennzeichnet, wie aus den Abb. 88 u. 89 ersichtlich ist.

Fazit:
Für einen plurizentrisch ausgewogenen DaF-Unterricht eignet sich also momentan nur die ADABA. DaZ in Deutschland kann sehr gut auch mit Tonspuren des DWDS und der DGD[28] arbeiten. Die DGD ist eine hervorragende Datenbank gesprochener Texte aus vielen Regionen v.a. Deutschlands, ihre Daten sind daher in erster Linie für die Dialektologie interessant.

Schlussbemerkung zur Studie 1:
Zum Schluss des phonetischen Teils der Korpusarbeit ist festzustellen, dass der Phonetik v.a. an Auslandsgermanistiken, die von der Tradition der russischen Fremdsprachendidaktik beeinflusst worden sind, viel Zeit gewidmet wird (vgl. Sorger 2012: 93). Trotzdem können die Studierenden am Ende ihres Studiums die Unterschiede in der Aussprache der drei Standardvarietäten kaum erkennen: Über Jahre hinweg habe ich Studenten/-innen der Brünner

[28] Ein sinnvoller Einsatz der DGD im DaZ muss wegen schlechter Qualität einiger Tonspuren und zum Teil starker Dialektfärbung einzelner Sprecher sehr gut überlegt werden.

Germanistik (Niveau C1/2) an drei kurzen Ansagen entscheiden lassen, aus welchem Land der Sprecher oder die Sprecherin kommt. Regional einzuordnen waren zwei Programmansagen (Karin Eppler, SAT1 und Chris Lohner, ORF) und ein kurzer Tonausschnitt aus der Themenübersicht der SRG-Nachrichtensendung „10 vor 10“ (Stefan Klapproth). Obwohl sich alle drei Aufnahmen durch „höchstes“ Deutsch, allerdings jeweils mit einem deutlich erkennbaren Regionalanschlag auszeichnen, konnten nur die wenigsten Studierenden (C1-Niveau) die richtige Varietät mit Sicherheit bezeichnen. Diesbezügliche mündliche Befragungen meiner Kolleg/-innen aus anderen tschechischen und slowakischen Universitäten bestätigen die Annahme, dass auf diesem Gebiet noch viel Arbeit (zumindest im DaF-Bereich in Tschechien und in der Slowakei) geleistet werden muss. Die Phonetik ist auf der plurizentrischen Seite offensichtlich noch unzureichend abgedeckt und ein gezieltes Sensibilisieren für diese Problematik wäre notwendig. Einfache Hinweise über das Ursprungland der Hörtexte in plurizentrisch halbwegs gut ausgebauten Lehrwerken reichen dazu offenbar nicht.

Studie 2: Suche nach dem „richtigen“ Schriftbild

Die Besonderheiten, Irregularitäten und Zweifelsfälle der deutschen Orthographie oder einfach auch die Tatsache, dass man sich nicht sicher ist, wie etwas üblicherweise geschrieben wird, erkundet man am schnellsten an Daten des größten Korpus für die deutsche geschriebene Sprache – DeReKo.

Fragestellung 1: *Welche Schreibweise überwiegt:* soft drink, soft Drink, Soft drink *oder* Softdrink*? Gibt es auch regionale Unterschiede?*

Da das Archiv der geschriebenen Sprache des **DeReKo** stilistisch und regional ausgewogen ist, wird jetzt hier recherchiert, um gegebenenfalls auch regionale Unterschiede beobachten zu können. Bei den Abfragen muss besonders die Einstellung von Groß-/Kleinschreibung in den Optionen beachtet werden.

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: **Suchmodalitäten:**
  - Groß- / Kleinschreibung beachten für 1. Zeichen — auswählen
  - Groß- / Kleinschreibung beachten für andere Zeichen — auswählen
  - **Expansionslisten:** — abwählen
  - Übernehmen
  - **Ergebnispräsentation:** Länderansicht — auswählen
- **Suchanfrage** (Eingabe ins Suchfeld)**:** **soft·drink** → Suchen
  dann: **soft·Drink** → Suchen
  dann: **Soft·drink** → Suchen
  dann: **Softdrink** → Suchen

Die Ergebnisse der einzelnen Abfragen sind in der folgenden Tabelle zusammengefasst:

Form	rel. Häuf. pMW[29]	Land
soft drink		
	0.0029	A
	0.0013	D
soft Drink		
	0	-
Soft Drink		
	0.0132	A
	0.0057	D
	0.0043	CH
Softdrink		
	0.0651	D
	0.0496	CH
	0.0396	A

Tab. 22: Schreibweise *Softdrink*/ *Soft Drink*/ *soft Drink*/ *soft drink*

Fazit:
Allgemein gesehen ist das Lexem *Softdrink* in der geschriebenen deutschen Sprache relativ selten. Dies erkennt man an der relativen Häufigkeit pro Mio. Wörter (mittlere Spalte).

[29] Relative Häufigkeit pro eine Million Worte.

Dennoch lässt sich Folgendes feststellen: diese Bezeichnung kommt in der ursprünglichen Form *soft drink* nur marginal, als *soft Drink* gar nicht vor. Alle Belege *soft drink* sind jeweils in einem englischen Satz eingebaut, z.B. *Take a soft drink.*
Die üblichste Schreibweise ist *Softdrink*, die Form *Soft Drink* kann gelegentlich, am ehesten in Österreich, vorkommen.
Ein ähnliches Problem stellen „Doppelnamen" (Allonyme) von Eigennamen dar.

Fragestellung 2: *Welche Form überwiegt:* Peking *oder* Beijing*? Gibt es auch regionale Unterschiede? Seit wann ist die Form* Beijing *belegt?*

Recherchiert wird wieder im **DeReKo**, angezeigt wird die Jahrzehnt- und Länderansicht.

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: **Suchmodalitäten:**
  - Groß- / Kleinschreibung beachten für 1. Zeichen — auswählen
  - Groß- / Kleinschreibung beachten für andere Zeichen — auswählen
  - **Expansionslisten anzeigen** — abwählen
  - mit Häufigkeiten — auswählen
  - **Sortierung:** nach Häufig. absteigend — auswählen
  - Übernehmen
  - **Ergebnispräsentation:** Länderansicht — auswählen
- **Suchanfrage** (Eingabe ins Suchfeld)**:** **Peking** → Suchen
  dann: **Beijing** → Suchen

Die Ergebnisse der Recherchen sind in der folgenden Tabelle (Tab. 23) zusammengefasst.

Form	rel. Häufigkeit pMW	von	bis	Land
Peking				
	26.38	1996	2013	CH
	21.11	1991	2013	A
	17.92	1949	2013	D
Beijing				
	1.340	1985	2013	D
	0.475	1996	2013	CH
	0.370	1991	2013	A

Tab. 23: Formen *Peking* und *Beijing* in einzelnen Ländern und Jahrzehnten

Aus der Ergebnistabelle ist ersichtlich, dass die Form *Beijing* sich in erster Linie in den Texten aus Deutschland eingenistet hat, und das bereits seit 1985. Die Belege aus dem Jahr 1985 sind aber eine Ausnahme: alle (insgesamt 8) sind in demselben Artikel der Wochenzeitung DIE ZEIT[30] zu finden. Bis 1991 ist *Beijing* nicht belegt. Ein signifikantes Vorkommen dieser Form erscheint erst ab 1995 (mit sehr wenigen Belegen auch zw. 1991 und 1994).
Beim Abrufen der Themenansicht bekommt man ein Bild über die Verteilung der Belege in einzelnen Textsorten. Die meisten Belege sind naturgemäß aus Zeitungstexten.

[30] Die Zeit, 05.04.1985, S. 09; Das größte Experiment der Geschichte. Quelle: DeReKo.

Über österreichische Texte vor 1991, bzw. schweizerische vor 1996 lassen sich keine Schlussfolgerungen ziehen, weil die ältesten Texte im Korpus eben aus den Jahren 1991 bzw. 1996 kommen.

Fazit:
Beijing kommt häufiger in bundesdeutschen Texten vor, heutzutage ist es eine gelegentliche Form auch in der Schweiz und in Österreich. *Peking* lebt auch heute noch in allen Varietäten.

Schlussbemerkungen zur Studie 2:
Analog zu den Recherchen können auch andere Lexeme mit schwankender Schreibweise untersucht werden. Es handelt sich um direkte Entlehnungen aus fremden Sprachen, deren Schreibweise zwar reglementiert wird, die tägliche Praxis kann davon jedoch (stark) abweichen. Dies gilt übrigens auch für zwanghafte Umsetzungen von Rechtschreibreformen. Korpora können Rückmeldungen ob der Akzeptanz der Rechtschreibreform liefern. Zum Beispiel die Wörter
Schluss und *dass* werden häufig auch nach 2005[31] *Schluß* bzw. *daß* geschrieben. Interessant wäre die Aufdeckung der Quellen, in denen die „alte" Schreibweise beibehalten wird.

Im DeReKo wurden Schriftbilder einiger Lexeme gefunden, wie sie in der Tab. 24 angeführt werden. die im Duden (2006) maximal als Dubletten oder gar (noch) nicht vorkommen:

DeReKo	**Duden (2006)**
E-Mail, e-mail, E-mail	*E-Mail*
Hot Dog, Hotdog, Hot-Dog, hot dog, Hot-dog, HotDog	*Hotdog, Hot Dog*
iPhone, IPhone, Iphone, iphone *i-Phone, I-Phone, I-phone, i-phone*	*keine Angabe*
Lavazza (ohne Kaffee), Lavazza-Kaffee, Lavazza-Caffè	*keine Angabe*

Tab. 24: Schriftbilder der Lexeme *E-Mail*, *Hot Dog*, *iPhone* und *Lavazza-Kaffee*

Die einzelnen Formen im DeReKo sind nach der Häufigkeit absteigend geordnet. Die Ergebnisse der Abfragen liefern auch andere interessante Informationen über die Lexeme, die hier nicht weiter ausgeführt werden können. Es sei nur bemerkt, dass *e-mail* die üblichere Schreibform bis etwa 2000 war und *iPhone*[32] offensichtlich die beliebteste Form für die Bezeichnung der entsprechenden Geräte ist, denn diese Schreibweise übertrifft alle anderen fast hundertfach.

Analog zur Fragestellung 2 ist auch der Wandel des Namens *Bombay* (seit 1996 offiziell) *Mumbai* zu beobachten. Im Korpus DeReKo kann man diesen Bruch deutlich beobachten: in deutschen und österreichischen Texten kommt *Bombay* vor 1996 (und auch danach) vor, *Mumbai* lediglich nach 1996 mit jährlich zunehmender Häufigkeit. (In schweizerischen Texten kann nur die zunehmende Häufigkeit von *Mumbai* nach 1996 verfolgt werden, denn die Texte aus der Schweiz sind im DeReKo nicht älter als 1996.)

Eine Studie, die zeigt, wie der inhaltliche Wandel mithilfe von Korpusdaten erkannt werden kann, ist in der Studie 12.

[31] Die letzte Rechtschreibreform ist seit dem 1. 8. 2005 verbindlich in Kraft.
[32] Nach Angaben einiger US-Amerikaner und Australier wird übrigens *iphone* langsam appellativiert (zur allgemeinen Bezeichnung wie *Tempo* für *Papiertaschentuch*) und zum Synonym von *mobile* (*phone*) oder *cell phone*.

Grammatik, Morphologie und Syntax

Der einzige „feste“ Punkt in der Sprache ist die Grammatik. Im modernen Fremdsprachenunterricht wird sie (aus verschiedenen Gründen) aus den Kontaktstunden in die individuelle Phase des Fremdsprachenerwerbs verlagert. Dennoch bleibt sie ein wichtiger Teil der Sprachkompetenz und muss trainiert werden.

Für die Erstellung (auch online) grammatikalischer Übungen gibt es kaum eine bessere Quelle als ein Korpus, weil die Beispiele aus dem sprachlichen „Alltag“ kommen, deswegen wirken (und sind) sie natürlich, man kann aus unzähligen Belegen den passenden wählen und darüber hinaus kann man – wenn mit einem Parallelkorpus gearbeitet wird – die Phänomene mit der Erstsprache der Lernenden vergleichen und ihnen so helfen, die Unterschiede besser zu verstehen.

Da die Rechercheergebnisse aus allen Korpora elektronisch in editierbarer Form dargestellt werden (doc-, rtf-, oder xls-Datei), können sie schnell in E-Learning-Übungen umgewandelt werden. Manchmal muss man natürlich die Belege modifizieren (kürzen) und/ oder solche auswählen, die dem sprachlichen Niveau der Lernenden entsprechen. Die mühevolle Suche nach passenden Beispielen einzelner grammatischer Phänomene in Büchern, Zeitungen, Zeitschriften oder im Internet fällt aber weg.

Im Folgenden werden exemplarisch einige korpusbasierte Übungsbeispiele zur deutschen Grammatik und Lexik vorgestellt, um zu demonstrieren, dass ein sicherer und zielführender Umgang mit Korpora viel Mühe und Zeit sparen kann und dass eine maßgeschneiderte Übung oder Prüfung mit authentischen Texten nicht besonders schwierig zu erstellen ist.

Studie 3: Grammatik auf einen Klick

Die folgenden Fragestellungen repräsentieren nur einen kleinen Ausschnitt aus dem Grammatikkomplex der deutschen Sprache. Sie beschränken sich auf einige wenige morphosyntaktische Phänomene, die absolut feste Regeln haben, dadurch können sie auch am PC trainiert und geprüft werden[33].

Fragestellung 1: *Wie erstellt man eine Übung zu Richtungsangaben?*

Gesucht wird nach dem Prädikatsverb *stellen* mit der Richtungsangabe *auf* (weiter auch *in, unter*, *über*, *hinter*): Satzmuster: *stellen etw. auf* (*in/unter*) *etw.*
Recherchiert wird im **InterCorp**, da dieses auch den Vergleich mit anderen Sprachen ermöglicht. Darüber hinaus ist die Abfrage im InterCorp einfacher als im DeReKo, obwohl es auf den ersten Blick vielleicht nicht so aussieht. Für diejenigen, die sich mit der Syntax der Abfrage vertraut gemacht haben, ist das Generieren der Abfrage simpel. Ihre Bedeutung wird anschließend erklärt.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de (ggfs. noch (eine) andere Parallelsprache(n) dazu)
- Query Type: CQL auswählen
- CQL (Suchfeldeingabe): **[lemma="stellen"] []{0,2} [word="auf"] []{0,2} [tag="NN"]** → Search

[33] Vgl. Grammatikkurse (B1) am Lehrstuhl für deutsche Sprache, Masaryk-Universität oder Káňa 2012.

Die Abfrage bedeutet: Suche nach dem Lemma *stellen*, Wortform *auf* und allen normalen Substantiven, dazwischen jeweils der Abstand von 2 Positionen.
Die Auswahl der Ergebnisse (in Konkordanzzeilen) präsentiert Abb. 90:

Quelle	links	KWIC	rechts
Skvorecky-	Sie	**stellte das Töpfchen auf ein Blechtablett**	und sagte : » Willst du ein Stück Napfkuchen ?
carroll-ale	Sie	**stellte sich also auf die Fußspitzen**	und guckte über den Rand des Pilzes…
simmel-ka	Ich stand auf . Wallace	**stellte die Lampe auf einen Tisch**	.
kundera-z	Ich stand dicht neben dem Pferd und	**stellte Wich auf die Fußspitzen**	, um mich mit meinen Lippen so nahe wie möglich
remarque-	Draußen war es frisch und klar . Ich	**stellte den Spirituskocher auf die Bank**	und suchte die Dose mit Kaffee . Meine Wirtin ,
fuks-pan_t	Er	**stellt das Essen auf den Tisch**	und fängt an zu essen .
Woodova-	Kazlah	**stellte das Fläschchen auf den Tisch**	neben dem Bett und richtete sich auf . Aller Augen
simmel-ka	Zu Mittag öffnete er Gulaschbüchsen ,	**stellte sie auf den Ofen**	und servierte sie dann mit Brot und Bier.
cook-toxin	Statt dessen	**stellte er sich auf die Zehenspitzen**	und sah sich nochmals nach seinem Patienten um.
konsalik-d	Sie setzte sich auf ,	**stellte die Füße auf den Boden**	und zündete sich eine Zigarette aus der Tabatiere ar

Abb. 90: Konkordanzen zur Abfrage: *stellen (…) auf (…)* (InterCorp_de)

Abfragen nach anderen Lokalpräpositionen mit dem Verb *stellen* sehen gleich aus, lediglich die Präposition in der Klammer **[word="auf"]** wird geändert: **[word="in"]** bzw. **[word="unter"]**. Entsprechend kann auch der Abstand geändert werden: **[]{0,2}**heißt Abstand von zwei Positionen, **[]{0,3}** heißt Abstand von drei Positionen.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query Type: CQL auswählen
- CQL (Suchfeldeingabe):
  [**lemma="stellen"] []{0,2} [word="in"] []{0,2} [tag="NN"]** → Search
  dann: **[lemma="stellen"] []{0,2} [word="unter"] []{0,2} [tag="NN"]** → Search

etc.

Zu jeder Abfrage bekommt man mehrere tausend Konkordanzzeilen. Es ist empfehlenswert, diese zu mischen (auf dem waagrechten Menübalken: Concordance → Shuffle), dadurch bekommt man Belege aus unterschiedlichen Quellen (siehe in der Abb. 90 die Spalte Quelle). Dann müssen nur diejenigen ausgewählt werden, die wirklich eine Lokalpräposition darstellen. Die meisten Belege repräsentieren nämlich Phraseme: *unter Verdacht stellen*, *unter (…) Willen stellen*, *in Frage stellen*, *in Rechnung stellen* und viele andere mehr.

Gut gewählte Konkordanzen können als Übungen zum Trainieren von gleich mehreren grammatikalischen Aspekten verwendet werden:
Ergänzungen von 1) Artikel; 2) (Personal-)Pronomen; 3) Präpositionen, wie die folgenden Beispiele zeigen:

1) Artikelergänzung:
Ergänzen Sie den passenden Artikel:

Ich packte die Wiege und **stellte** sie (…) **neben ___** Kopfende meines Bettes. (das)
Die Herren **stellen** sich **in ___** Ecke und politisieren. (eine)
Kalmat **stellte** sich ihnen **in ___** Weg. (den)
Spann deinen Schirm auf und **stell** ihn **in ___** Ausguß, da kann er trocknen. (den)

2) Ergänzung von Pronomen
Ergänzen Sie das passende Personal- oder Possessivpronomen:

Eines Abends **stellte** er sich **unter ___** Fenster, als sie gerade mit ein paar Freundinnen Tee trank. (ihr)
Stellen Sie sich **hinter ____,** Rücken an Rücken! (mich, ihn, sie)

3) Ergänzung von Präpositionen:

Ergänzen Sie die passende Präposition:

Ich packte die Wiege und **stellte** sie (…) **_____ das Kopfende** meines Bettes. (neben, an, (auf))

Die Herren **stellen** sich **___ eine Ecke** und politisieren. (in)

Kalmat **stellte** sich ihnen **___ den Weg**. (in)

Spann deinen Schirm auf und **stell** ihn **___ den Ausguß**, da kann er trocknen. (in)

Eines Abends **stellte** er sich **____ ihr Fenster**, als sie gerade mit ein paar Freundinnen Tee trank. (unter)

Stellen Sie sich **____ mich**, Rücken an Rücken! (hinter)

Die Leerstellen können im elektronischen Übungsportal durch ein Eingabefeld oder Auswahlfensterchen ersetzt werden.

Stellen Sie sich **_____** mich, Rücken an Rücken!

Stellen Sie sich **hinter** mich, Rücken an Rücken!
zu
bei

Fragestellung 2: *Wie erstellt man eine Übung zum Gebrauch von Dativ und Akkusativ nach einer Präposition?*

Die Recherche kann wieder im **InterCorp** durchgeführt werden, wegen Absenz einiger Sprachen im InterCorp wird auch im **OPUS-Corpus** recherchiert.

Recherche im InterCorp

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de (ggfs. noch (eine) andere Parallelsprache(n) dazu)
- Query Type: CQL
- CQL (Suchfeldeingabe): **[word="auf"] []{0,2} [tag="NN"]** → Search

Die Abfrage bedeutet: Suche nach der Wortform *auf* und allen normalen Nomina im Abstand von 2 Positionen.

Nach dem die Konkordanzen erschienen sind:

- Frequency → Node forms.

Als Ergebnis bekommt man die Statistik einzelner Kombinationen nach der absoluten Häufigkeit des Vorkommens im Korpus. Die 50 häufigsten Verbindungen im InterCorp sind auf der folgenden Seite in der Abb. 91 aufgelistet.

Frequenzliste der Formen auf + Substantiv (gekürzt)

Total: 42612 (853 pages)							
		word	**Freq**			**word**	**Freq**
1.	p/ n	**auf den Tisch**	1.017	26.	p/ n	**auf Erden**	195
2.	p/ n	auf der Straße	860	27.	p/ n	auf den ersten Blick	192
3.	p/ n	**auf den Boden**	794	28.	p/ n	**auf den Beinen**	190
4.	p/ n	**auf dem Boden**	761	29.	p/ n	auf die Idee	182
5.	p/ n	auf der Welt	582	30.	p/ n	auf der Treppe	180
6.	p/ n	auf die Straße	549	31.	p/ n	**auf die Beine**	176
7.	p/ n	**auf dem Weg**	548	32.	p/ n	auf dem Fußboden	174
8.	p/ n	**auf den Weg**	504	33.	p/ n	auf die Brust	167
9.	p/ n	**auf dem Tisch**	489	34.	p/ n	auf der Brücke	167
10.	p/ n	auf die Schulter	483	35.	p/ n	auf die Stirn	165
11.	p/ n	auf dem Rücken	473	36.	p/ n	auf der Insel	158
12.	p/ n	**auf der anderen Seite**	472	37.	p/ n	auf dem Hof	156
13.	p/ n	**auf der Erde**	417	38.	p/ n	auf Grund	155
14.	p/ n	auf der Stelle	401	39.	p/ n	auf dem Gang	154
15.	p/ n	auf den Rücken	370	40.	p/ n	auf dieser Welt	153
16.	p/ n	auf diese Weise	342	41.	p/ n	auf . „	153
17.	p/ n	**auf dem Kopf**	331	42.	p/ n	auf einen Stuhl	152
18.	p/ n	auf die Uhr	325	43.	p/ n	auf die andere Seite	150
19.	p/ n	auf den Kopf	309	44.	p/ n	auf dem Platz	150
20.	p/ n	auf der Suche	279	45.	p/ n	auf dem Friedhof	143
21.	p/ n	auf keinen Fall	236	46.	p/ n	**auf die Seite**	140
22.	p/ n	**auf die Erde**	236	47.	p/ n	auf den Gedanken	140
23.	p/ n	auf jeden Fall	233	48.	p/ n	auf den Mund	139
24.	p/ n	auf dem Bett	228	49.	p/ n	auf der Schwelle	136
25.	p/ n	auf den Knien	215	50.	p/ n	auf den Grund	135

Abb. 91: Die 50 häufigsten Kombinationen von *auf* + Substantiv im Abstand von 2 Positionen (InterCorp_de)

Zu beobachten ist ein interessantes Phänomen: einige Verbindungen mit demselben Substantiv (in Fettschrift) weisen (sehr grob gesehen) ähnliche Frequenz auf: *auf den/dem Tisch, auf den/dem Boden, auf den/dem Weg*. Dieses Phänomen muss noch mit Einbeziehung der dazugehörigen Verben untersucht werden. Dies erfolgt am einfachsten über eine Kollokationsanalyse.

Als Beispiel werden die Verbindungen *auf den/dem Tisch* gewählt:
Beim Klicken auf p (positiver Filter) bei der Akkusativverbindung *auf den Tisch* öffnen sich die (parallelen) Konkordanzen. Jetzt kann man die Kollokationsanalyse durchführen:

- p (positiver Filter) anklicken

Nach dem die Konkordanzen erschienen sind:

- Collocations → Custom… (Standardeinstellung) → Make Canditate List

Denselben Vorgang unternimmt man mit der Dativverbindung.

Die Berechnung der Kollokationen zeigt folgende signifikante Kollokationspartner:

Akkusativverbindung: *auf den Tisch*
stellen, bringen, schlagen, hauen …

Dativverbindung: *auf dem Tisch*
liegen, stehen, ausbreiten, sitzen, haben …

Die Analyse zeigt hochfrequentierte Chunks in der geschriebenen Sprache. Sie sind auch in der gesprochenen Sprache üblich – dies belegen auch Beispiele in den Abb. 92 und 93 in der direkten Rede. Aus diesen Gründen sollte solchen Verbindungen im Unterricht Aufmerksamkeit gewidmet werden.

Er **stellte** sein Glas	auf den Tisch	und sagte:
Jonas **brachte** Kürbisse	auf den Tisch	, …
Der Erste Offizier **schlägt**	auf den Tisch	und erklärt ihm …
Da **hab**' ich	auf den Tisch	**gehauen** und staunte zugleich über mich: (…)

Abb. 92: Auswahl aus Konkordanzzeilen zur Abfrage *auf den Tisch* (InterCorp_de)

„Die Blumen für dich **liegen** im Wohnzimmer	auf dem Tisch	“, sagte der Vater.
Wie viele Weingläser **stehen**	auf dem Tisch	?
… und nahm eine (…) Karte heraus, die er	auf dem Tisch	ausbreitete.
Der Feldkurat **saß** recht bequem	auf dem Tisch	und drehte sich eine Zigarette.
„Wir **haben** nichts	auf dem Tisch	(…)“, erklärt der Senatspräsident …

Abb. 93: Auswahl aus Konkordanzzeilen zur Abfrage *auf dem Tisch* (InterCorp_de)

Ausgewählte Konkordanzzeilen (Sätze oder Textpassagen) bieten wieder mehrere Möglichkeiten an, wie sie in Übungen umgewandelt werden können - Artikelergänzung oder Ergänzung von Verben, um nur zwei zu nennen:

1) Artikelergänzung:
Ergänzen Sie den passenden Artikel:
Er **stellte** sein Glas **auf ___ Tisch** und sagte: (den)
… zog eine Schublade auf und nahm eine zusammengerollte Karte heraus, die er **auf ___ Tisch ausbreitete**. (dem)
etc.

2) Ergänzung von Verben
Ergänzen Sie das passende Verb:
Der Feldkurat **___** recht bequem **auf dem Tisch** und drehte sich eine Zigarette. (saß)
Da hab' ich **auf den Tisch _____** und staunte zugleich über mich: Mein Gott, so energisch bin ich doch sonst nicht. (gehauen)
etc.

Die Leerstellen können in einem Übungsportal durch Einsetzfelder oder Multiple-Choice-Auswahl ersetzt werden.
Falls am Anfang der Abfrage eine Parallelsprache ausgewählt wurde, werden zu den Konkordanzen die Parallelpassagen immer automatisch abgerufen.

Recherche im OPUS-Corpus

Ist die gewünschte Sprache im InterCorp (noch) nicht vertreten, kann man zum **OPUS-Corpus** ausweichen, wo auch einige außereuropäische Sprachen recherchierbar sind. InterCorp verfügt d.Z. über keine türkische Parallele, dennoch können Lernende mit dieser Erstsprache bzw. deutschsprachige Türkischlernende auch das hier betrachtete Phänomen (*auf* + Tisch im Dativ oder Akkusativ) mit dem Deutschen vergleichen.

- **OPUS-Corpus** → Search & Browse → OpenSubtitles search interface
- **lang = de** (ganz rechts)
- CQP Query (Suchfeldeingabe):
  im Suchfeld **"viol.*"** löschen, statt dessen: **"auf"·".*"·"Tisch"** eingeben.
- Display: context = sentence (zur Auswahl 5 Wörter, 2 Sätze etc.) auswählen
- Alignments: **tr** (Türkisch) auswählen → Run Query

Beim Eingeben der Suchanfrage Achtung auf den richtigen Abstand und auf die Anführungszeichen!

Die Ergebnisse erscheinen in Form einer Tabelle, wie die Abb. 94 zeigt:

de	**tr**
Pack die Scheine wieder **auf den Tisch**, ich bin dran mit Geben.	- Senin için sevindim ama burası benim garajım evlat !
Alles **auf diesem Tisch** wurde hier geerntet.	Masadaki her şey burada yetiştirildi .
Du schüttest ihn runter und haust das leere Glas **auf den Tisch**.	Önce fondip yapıp bardağı masaya vuracaksın , tamam mı ?
Und die Leute , die **auf einen Tisch** warten ?	- Masa bekleyenleri nereye alacağım ?
Leg die Ellenbogen **auf den Tisch**, beug dich über den Brief und lies ihn vor.	Dirseklerini masaya koy … eğil … yüzünü mektuba yaklaştır ve yüksek sesle oku .
Leg die Hände **auf den Tisch**.	Avuçların aşağıda kalacak şekilde , ellerini masaya koy .
Ich habe sie Ihnen **auf den Tisch** gelegt.	- İstersiniz diye düşündüm , masanızın üzerine koydum

Abb. 94: Konkordanzen zur Abfrage *auf … Tisch* (OPUS-Corpus, de - tr)

Aufgaben und Übungen können analog zu den oben angeführten erstellt werden, dabei *kann* (!) die parallele Sprache die Übung entlasten:

> Alles **auf diesem Tisch** wurde hier geerntet.
> **Masadaki** her şey burada yetiştirildi.

Die Betonung von *können* ist hier wichtig. Man muss immer davon ausgehen, dass es sich um parallele Texte handelt, die „funktional-äquivalent" (Knittlová 2009: 7), nicht struktur-äquivalent sind. Die parallelen Belege sollen entlasten, nicht Fragezeichen über die lexikalische und grammatikalische Äquivalenz aufkommen lassen. Deswegen müssen die Belege zur Übung von grammatikalischen Erscheinungen sorgfältig ausgesucht werden. Darüber hinaus muss man auch bedenken, dass es sich in den Korpustexten oft um keine Übersetzungen zwischen den zwei ausgewählten Sprachen handelt, sondern, dass die Texte vielleicht aus einer dritten Sprache übersetzt wurden, wie es auch hier in der Abb. 94 der Fall ist. Die Texte sind aus dem Englischen ins Deutsche und ins Türkische übersetzt worden, es handelt sich um Untertitel zu amerikanischen Filmen. Dadurch dürfen die Entsprechungen nicht nur formal, sondern auch semantisch etwas weiter voneinander liegen.

Fragestellung 3: *Wie erstellt man eine Übung zur Deklination von Artikel und Adjektiv nach einer Präposition für Lernende mit ukrainischer oder türkischer Erstsprache?*

Wenn der Kontrast zu einer anderen Sprache benötigt wird, ist für die Übungserstellung das **InterCorp** das geeignete Instrument.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de (ggfs. noch (eine) andere Parallelsprache(n) dazu)
- Query Type: CQL
- CQL (Suchfeldeingabe): **[tag="APPR"] [tag="ART"] [tag="ADJA"] [tag="NN"]** → Search

Diese Suchanfrage ruft aus dem Korpus alle Wortketten ab, in denen Präposition – Artikel – Adjektiv – Nomen (ohne Eigennamen) nacheinander vorkommen:

der (…) gekrümmt in der Ecke	**unter einer kleinen Palme**	stand.
Hermine hatte das Bild dann	**mit einem kleinen Zaubertrick**	(…) zum Leuchten gebracht.
	In der kleinen Stadt	wurde es bald allgemein bekannt …
Dann trocknet er sich die Stirn	**mit einem rotkariertem Taschentuch**	.
Zuerst wollte er	**mit dem unteren Teil**	seines Körpers aus dem Bett…
… und sie hielten sich	**in den tiefsten Löchern**	verborgen.
…, so daß ich	**wegen einem dummen Faß**	Benzin nicht auslernen hab können.

Abb. 95: Konkordanzen zur Abfrage: Präposition - Artikel - Adjektiv - (normales) Nomen (InterCorp)

Die Konkordanzen können in Ergänzungsübungen umgewandelt werden:

… der linkisch gekrümmt in der Ecke unter **ein__ klein__** Palme stand.
das Bild mit **ein__ klein__** Zaubertrick zum Leuchten bringen
In **d__ klein__** Stadt wurde es bald allgemein bekannt, dass …

Und natürlich kann man nach Bedarf entsprechende Parallelen in einer anderen Sprache zur Verfügung stellen (hier Ukrainisch):

… der unter ein__ klein__ Palme stand
… стояв , втягнувши голову в плечі , **під карликовою пальмою** .
--
Hermine hatte das Bild dann **mit ein__ klein__ Zaubertrick** in verschiedenen Farben zum Leuchten gebracht.
А Герміона **з допомогою одного заклинання** зробила так , що фарба на плакаті стала мінитися різними барвами .
--
In d__ klein__ Stadt wurde es bald allgemein bekannt, dass …
У малесенькому містечку швидко поширилася чутка , що …

Fazit und Schlussbemerkung zur Studie 3:

Mit einer klug gestellten Abfrage können viele grammatikalische Phänomene abgefragt werden. Einige Beispiele wurden hier ausführlicher angeführt. Weitere Abfragen ergeben sich aus den kommenden Studien, viele bleiben jedoch offen. Es liegt an den Korpusbenutzern und -benutzerinnen selbst, wie kreativ sie mit Korpusmanagern umgehen (können). Aus den Aufgaben ist jedoch ersichtlich, dass es sinnvoll ist, die morphologische Annotation (Tags) zumindest teilweise zu beherrschen.

Abschließend muss noch betont werden, dass die Möglichkeiten, die Korpora für die einfache, aber auch kontrastive Grammatikvermittlung anbieten, fast berauschend ist. Diese Tatsache

darf aber natürlich nicht zur Wiederkehr der Grammatik-Übersetzungsmethode führen und verführen.

Morphologie und Syntax

Die Grammatik deckt sich teilweise mit den linguistischen Disziplinen Morphologie und Syntax. Bei der systematischen Analyse der Korpusdaten überschneiden sich auch diese beiden Disziplinen in vielen Punkten. Die Morphologie befasst sich mit den kleinsten sprachlichen Einheiten, die eine „Bedeutung[34]“ haben. Grammatikalische Morpheme sind „Biegungselemente“, die einzelne morphologische Kategorien (grammatikalisches Geschlecht, Zahl, Steigerungsform, Tempusform u.a.) repräsentieren.

Die Syntax befasst sich wiederum damit, wie die einzelnen Elemente der Sprache (morphologisch richtig) in sinnvollen Sätzen gebildet werden. Da aber die morphologischen Kategorien unterschiedlich realisiert werden – *blieb* (ist eine rein morphologisch gebildete Tempusform) vs. *bin geblieben* (ist eine morphosyntaktisch gebildete Tempusform) - scheint es sinnvoll, die beiden Ebenen zu verbinden. Dieser Schritt hat mehrere Gründe: 1) Die Grenze zwischen der Morphologie und Syntax ist fließend (vgl. Skalička 1957). Betrachtet man rigide die Morphologie als eine Disziplin, die sich mit sprachlichen Elementen bis zur Wortgrenze befasst, dann ist Deutsch morphologisch ziemlich arm (Englisch und Spanisch hätten demnach fast keine morphologischen Formen); 2) Würde man typologisch unterschiedliche morphologische Systeme von Sprachen strikt vergleichen, dann gäbe es viele Null-Äquivalente (Divergenzen), weil es grammatikalische Kategorien gibt, die sich in einer Sprache morphologisch demonstrieren, während sie in einer anderen Sprache syntaktisch gebildet werden; 3) Im modernen Fremdsprachenunterricht spielt die Aufteilung in Morphologie und Syntax eher eine geringere Rolle. Die „grammatikalische[35] Richtigkeit“ tritt zugunsten der kommunikativen Kompetenzen in den Hintergrund. Diese Kompetenzen zeigen sich in erster Linie auf der lexikalisch-semantischen und pragmatischen Ebene. Die Lernenden werden im heutigen Unterricht oft aufgefordert, selbst morphologische und syntaktische Regularitäten aufzudecken und so ihre produktiven Fertigkeiten auf Richtigkeit zu prüfen. Dazu können ihnen auch Korpora behilflich sein.

Im Folgenden werden einige wenige Strukturen anhand von Korpusdaten exemplarisch erforscht. Die Aufgaben zu einzelnen morphosyntaktischen Phänomenen können in Grammatikübungen umgewandelt werden.

Studie 4: Entdeckung der Flexionsformen

Wie die einzelnen Flexionsformen/ Biegungsformen eines beliebigen deutschen Wortes aussehen, können erfahrene Lehrer/-innen ohne Recherchen aufzählen. Lernende können sie in (guten) Wörterbüchern nachschlagen, aber es gibt auch Dubletten/ Doppelformen, bei denen man sich oft nicht sicher ist, wann welche gebraucht wird. In Wörterbüchern ist wenig Platz für alle plausiblen Beispiele in aufschlussreichem Kontext. Auch Informationen darüber, ob ein Wort, das mehrere syntaktisch-semantische Funktionen hat, mehrheitlich die eine oder

[34] Die Bedeutung kann grammatikalischer oder lexikalischer Natur sein. Da die grammatikalische „Bedeutung“ nicht ganz dem üblichen Sinn des Wortes „Bedeutung“ entspricht, ist dieses hier in Anführungszeichen gesetzt.

[35] Die Grammatik deckt sich nur teilweise mit den Begriffen Morphologie und Syntax. Dazu auch Fachlexikon (2010: 106-107).

die andere Funktion vertritt, werden selten angegeben. Eine Korpusstudie zu diesem Thema dauert zwar länger als das Nachschlagen in einem Wörterbuch oder einer Grammatik, liefert aber ziemlich sichere und realitätsnahe Ergebnisse.

Fragestellung 1. *Welche Formen hat das Verb* werden*? Kommt es in der Sprache öfter als Vollverb oder als Hilfsverb/ Auxiliar vor?*

Diese Aufgabe lässt sich einfach auch „manuell" an einem beliebigen Text lösen, denn das Verb *werden* ist im Deutschen (nach *sein*) das zweithäufigste Verb überhaupt (vgl. DeReWo 2009). Schneller und effektiver geht die Recherche in einem Korpus. Das Ergebnis ist auch objektiver, wenn mehrere unterschiedliche Texte einbezogen werden.

Für die Recherche eignet sich jedes getaggte Korpus. Das Tagging braucht man zur automatischen Analyse der syntaktisch-semantischen Funktionen des Verbs (*werden* als Auxiliar oder als Vollverb). Einfach und schnell geht es mit **InterCorp**.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query Type: Lemma auswählen
- Lemma (Suchfeldeingabe): **werden** → Search

Nach dem die Konkordanzen erschienen sind:

- Frequency → Node forms.

dann

- Frequency → Node tags.

Die Ergebnisse erscheinen in automatisch erstellten Tabellen (Abb. 96, 97). In der Abb. 96 werden die Formen (Spalte word) des Verbs nach der Häufigkeit (Freq.) angeführt, in der Abb. 97 sind seine Funktionen (Spalte tag) auch nach der Frequenz geordnet zu sehen. Um die syntaktisch-semantischen Funktionen richtig interpretieren zu können, muss man die Kürzel des Tagsets entschlüsseln können (siehe dazu Kap. 8.2).

Total: 39 (1 pages)

		word	Freq	Freq [%]
1.	p/ n	werden	310,128	40.6
2.	p/ n	wird	201,263	26.4
3.	p/ n	wurde	82,681	10.8
4.	p/ n	wurden	49,964	6.5
5.	p/ n	würde	37,813	5.0
6.	p/ n	worden	30,638	4.0
7.	p/ n	werde	14,520	1.9
8.	p/ n	würden	13,966	1.8
9.	p/ n	geworden	8,639	1.1
10.	p/ n	wirst	2,619	0.3

Abb. 96: Frequenzliste der Formen des Verbs *werden*

Total: 7 (1 page)

		tag	Freq	Freq [%]
1.	p/ n	VAFIN	562,707	73.7
2.	p/ n	VAINF	160,726	21.0
3.	p/ n	VAPP	39,085	5.1
4.	p/ n	VVFIN	894	0.1
5.	p/ n	VVPP	192	0.0
6.	p/ n	VVINF	95	0.0
7.	p/ n	VAIMP	3	0.0

Abb. 97: Frequenzliste der Funktionen des Verbs *werden*

Fazit:
In der Abb. 96 sind die ersten 10 üblichsten Konjugationsformen des Verbs *werden* angeführt. Markant ist die extrem hohe Frequenz der „Grundform" *werden* (40% aller Belege). Hier spielt natürlich auch die Homonymie der Formen eine wichtige Rolle, denn der Infinitiv und die 1. u. 3. Person Pl. Präsens sind formal ident. Wie die Verteilung unter diesen Kategorien aussieht, lässt sich grob durch weitere Schritte erkennen: positiver Filter p → Node Tags. Dies ergibt, dass es sich ausschließlich um Hilfsverben[36] handelt.

Aus der Abb. 97 ist ersichtlich, dass das Verb in erster Linie als ein Hilfsverb fungiert: über 73 % (fast ¾) aller Vorkommen des Verbs *werden* entfallen auf seine finite Form[37] (VAFIN = Auxiliarverb, finite Form) als Auxiliar. Weitere 21% der Belege bilden infinitive Formen des Auxiliars, insgesamt sind also fast 95% der Wörter *werden* Hilfsverben. Seine Funktion als Vollverb (Kürzel beginnend mit VV) ist vergleichsmäßig marginal, trotzdem beträgt sie über 1.000 Belege im Korpus. Diese Belege müssten jedoch noch sorgfältig aussortiert werden. Mehr als die Hälfte von ihnen wurde vom automatischen Tagger dort falsch erkannt (und annotiert), wo das Verb *werden* in einer für ein Auxiliar vergleichsweise eher untypischen Position steht (z.B. am Satzanfang) und das Vollverb sehr weit davon, weshalb es vom Tagger unerkannt blieb. Es ist darauf hinzuweisen, dass die automatische Annotation sich sehr stark auf die prototypische Verbindung der Wortform und ihrer Funktion stützt – in diesem Fall wird beim Tagging zu jeder Form von *werden* in demselben Satz nach einem zweiten Verb im Infinitiv oder Partizip Perfekt gesucht. Wird der Infinitiv, das Partizip Perfekt oder gar die Satzgrenze (vom ersten Großbuchstaben bis zum ersten Satzendzeichen) falsch erkannt, entsteht ein Annotationsfehler. Damit ist immer zu rechnen. Je eindeutiger die Funktion einer Wortform in der Sprache ist, desto niedriger ist die Wahrscheinlichkeit, dass die Wortform einen falschen Tag bekommt.

Zum Schluss dieser Recherche muss festgehalten werden, dass das präsentierte Ergebnis über die syntaktisch-semantische Verteilung der Funktionen des Verbs *werden* zwar naheliegend ist und dem linguistischen Bauernvernunft vollkommen entspricht, durch viele Korpusdaten ist diese These nun aber auch mehrfach (an verschiedenen Texten) bewiesen. Beispiele können sofort durch Klicken auf p (in der 2. Spalte) präsentiert werden.

Fragestellung 2. *Welche Formen hat das Verb* backen*?*

Die Konkurrenz der schwachen und starken Deklinations- und Konjugationsformen ist ein Evergreen in der deutschen Morphologie sowie eine Qual im traditionellen Unterricht. In der Sprache gibt es aber (pragmatisch gesehen) kaum etwas Unwichtigeres für eine erfolgreiche Kommunikation. Trotzdem wird nach der typischen Verwendung der sich konkurrierenden Formen oft gesucht. In diesem Falle können sie im **InterCorp** leicht gefunden werden. Die Abfrage ist gleich wie die in der vorherigen Fragestellung.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query Type: Lemma auswählen
- Lemma (Suchfeldeingabe): **backen** → Search

Nach dem die Konkordanzen erschienen sind:

- Frequency → Node forms.

[36] In der nächsten Phase des Korpustaggings wird auch das Abrufen anderer morphologischen Kategorien möglich (Person, Numerus, Tempus, Modus, Genus Verbi).
[37] Die häufigste finite Form ist die 3. Pers. Singular.

Die häufigsten Formen des Verbs *backen* sind aus der Abb. 98 ersichtlich.

Frequenzliste der Formen: Lemma *backen*

		word	Freq	Freq [%]				word	Freq	Freq [%]
1.	p/ n	gebacken	88	36.1		7.	p/ n	backe	11	4.5
2.	p/ n	backen	49	20.1		8.	p/ n	buk	10	4.1
3.	p/ n	Back	25	10.2		9.	p/ n	buken	8	3.3
4.	p/ n	back	24	9.8		10.	p/ n	backten	3	1.2
5.	p/ n	bäckt	11	4.5		11.	p/ n	backt	3	1.2
6.	p/ n	backte	11	4.5		12.	p/ n	gebacken	1	0.4

Abb. 98: Formen zum Lemma *backen* (InterCorp)

Die einzelnen Formen wurden weiter in einer Seminararbeit (Mair 2013) nach Texten und Zeit ihrer Entstehung untersucht und noch mit Daten des DeReKo verglichen. Die Ergebnisse zeigten deutlich, dass die schwachen Formen *backte*/ *gebackt* „im Vormarsch" sind. Die Form *backte* kommt allgemein sogar wesentlich häufiger vor als die starke Form *buk*. Sie kommt signifikant häufiger in bundesdeutschen und schweizerischen Texten vor, und zwar sowohl in der direkten als auch in der indirekten Rede. Die starke Form *buk* ist über die Länder gleichmäßig verteilt, markant ist jedoch der historische Touch dieser Form, denn in modernen deutschen Texten kommt sie vor allem in Verbindung mit Till Eugenspiegels Streich vor, wie die automatisch berechneten syntagmatischen Muster in der Abb. 99 zeigen (grau unterlegt), während die Form *backte* in Verbindungen mit semantisch relevanten Lexemen gleichmäßig verteilt ist. (Für den allgemeinen Sprachgebrauch „irrelevante" Verbindungen: homonyme Formen aus slawischen Sprachen (Abb. 100, Zeilen 1, 3, 4, 6) und linguistische Behandlungen der schwachen und starken Konkurrenzformen (Abb. 99, Zeile 3; Abb. 100, Zeilen 3, 9, 21) sind durchgestrichen.)

Kookkurrenzen zu *buk* (gekürzt)

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
1	7	7	107	~~**Skradinski**~~	~~100% Skradinski buk~~
2	16	9	94	**Meerkatzen** Eulen	77% er die Eulen und Meerkatzen buk
	18	2		Meerkatzen	100% Meerkatzen [...] buk
3	27	9	81	~~**backte**~~	~~44% buk ... backte~~
4	30	3	41	~~**slaw**~~	~~100% slaw [...] buk ... Buche~~
5	31	1	34	**Brot** Eulen	100% buk Eulen ... Brot
	62	31		Brot	58% buk [...] Brot
6	63	1	30	~~**slap**~~	~~100% slap ... buk~~
7	64	1	28	**Eulen**	100% buk ... Eulen
8	70	6	25	**Reibekuchen**	100% buk [... leckere] Reibekuchen und ...
9	80	10	23	**Waffeln**	90% buk [leckere|frische] Waffeln
10	87	7	21	**Backofen**	57% im Backofen [...] buk
11	102	15	21	**Brötchen**	80% buk [... kleine|kleinere] Brötchen
12	109	7	20	**kochte**	71% kochte und buk
13	115	6	19	**Brote**	83% buk ... Brote
14	122	7	17	**Kekse**	85% buk [...] Kekse
15	133	11	15	**Bäcker**	72% Ein findiger Bäcker [...] buk daraufhin Dreck
17	139	3	8	**Torten**	66% buk ... Torten
18	142	3	7	**Pfannkuchen**	100% buk [...] Pfannkuchen
19	143	1	7	**Eulenspiegel**	100% buk ... Eulenspiegel
20	146	3	6	**Backstube**	100% Backstube ... und buk
21	150	4	6	**Bäckermeister**	50% buk Bäckermeister

Abb. 99: Kookkurrenzen und syntagmatische Muster zur Abfrage: Wortform *buk* (DeReKo)

Kookkurrenzen zu *backte* (gekürzt)

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
1	47	47	197	**Waffeln**	85% backte [...] Waffeln und
2	72	25	137	**kochte**	76% kochte [und] backte
3	81	9	81	~~**buk**~~	~~44% backte ... buk~~
4	112	31	67	**Brötchen**	74% backte [... kleine] Brötchen und
5	135	23	50	**Plätzchen**	56% backte [...] Plätzchen und
6	138	3	41	**Konditor-Meisterin**	100% Konditor-Meisterin backte auf dem
7	153	15	38	**leckere**	86% backte [...] leckere
8	159	6	33	**Weihnachtsplätzchen**	50% backte der ... Weihnachtsplätzchen
9	162	3	33	~~**gebackt**~~	~~100% - backte - gebackt~~
10	213	51	27	**Kuchen**	76% backte [... einen] Kuchen für|und
12	251	32	21	**Brot**	71% und backte [...] Brot
14	256	3	15	**Meerkatzen**	66% backte ... Eulen und Meerkatzen
15	262	6	13	**Kekse**	83% backte ... Kekse
16	265	3	12	**Streuselkuchen**	100% backte [leckeren] Streuselkuchen
17	268	3	11	**Reibekuchen**	100% backte [...] Reibekuchen
18	274	6	10	**Bäckermeister**	66% backte [...] Bäckermeister
19	276	2	9	**Crêpes**	100% backte [...] Crêpes
20	279	3	9	**Kartoffelpuffer**	100% backte die ... Kartoffelpuffer
21	281	2	9	~~**backt**~~	~~50% backte ... backt~~
23	285	3	8	**Konditor**	66% backte [...] Konditor
24	288	3	8	**Muffins**	100% backte [...] Muffins
25	292	4	7	**Brote**	50% backte ... Brote
26	296	4	6	**Torten**	75% backte die ... Torten

Abb. 100: Kookkurrenzen und syntagmatische Muster zur Abfrage: Wortform *backte* (DeReKo)

Die Form des Partizips II von *backen* lautet in den meisten deutschsprachigen Texten *gebacken*. Nur gelegentlich kommt auch die schwache Form *gebackt* vor. Diese ist jedoch ausschließlich in bundesdeutschen Texten und meistens nur in der Wiedergabe der direkten Rede zu finden, wie die Konkordanzen in der Abb. 101 zeigen.

Konkordanzen zu *gebackt* (Auswahl)

RHZ01	kurze Zeit "gehen", bevor die Waffeln **gebackt** werden können.
RHZ01	Beitrag geleistet. Plätzchen wurden **gebackt**, kleine Geschenke
RHZ06	Erst wurde eifrig **gebackt**, und dann verzehrt
RHZ12	wollen. „Da haben wir kleine Brötchen **gebackt**. Wir waren uns
RHZ12	meiner Mutter: „Ich honn noch enn Sort **gebackt**", und nannte noch
WDD11	da 2 Euro." "Die Faasekichelcha genn **gebackt**."Rheinfränkisch:
WDD11	gebackt." Rheinfränkisch: "... werre **gebackt**." --Speifensender
WDD11	dies da ich die jahrelang als Bäcker **gebackt** habe. Heute gibt

Abb. 101: Konkordanzen zur Abfrage: Wortform *gebackt* (DeReKo)

Fazit:
In Anbetracht der Tatsache, dass die Recherchen in Texten der geschriebenen Sprache durchgeführt wurden (die gesprochene Sprache ist in der Entwicklung der geschriebenen voran), könnte eine Tendenz zu den schwachen Formen betrachten werden. Die Standardform des Imperfektums von backen ist *backte*, die Standardform des Partizips II ist immer noch gebacken, es ist vielleicht nur eine Frage der Zeit, dass auch diese mehrheitlich „schwach" gebildet wird. Eine regionale Differenzierung ist dabei nicht ausgeschlossen.

Schlussbemerkung zur Studie 4:
Die Konjugations- bzw. Deklinationsformen lassen sich in den meisten Korpora über die Funktion Lemma abrufen. Im DeReKo muss man die Einstellungen der Optionen beachten (Optionen → Lemmatisierung, es sollen nur Flexionsformen ausgewählt bleiben.) Man muss bedenken, dass die Korpora automatisch lemmatisiert sind, es können also viele Ballastformen mitabgerufen werden (z.B. zum Lemma des Verbs *fließen* sind es auch das Substantiv *Floss*/ *Floß*). Dennoch kann man sich darauf verlassen, dass gut lemmatisierte Korpora auch suppletive Formen sicher erkennen (*gut*, *besser*, *beste*) können.

Die Erweiterung von morphologischen (Deklinations- bzw. Konjugations-)Formen stellen Wortbildungsformen dar. Auch diese lassen sich gut erschließen, wie die folgende Studie zeigt.

Studie 5: Ermittlung einer Wortfamilie

Die lexikalische Kompetenz der Lernenden wird dadurch gefördert, dass auf formale und semantische Zusammenhänge aufmerksam gemacht wird. Die formalen Zusammenhänge im Lexikon sind an einzelnen Gliedern einer Wortfamilie deutlich zu sehen. Die Glieder einer Wortfamilie lassen sich schnell abrufen, dabei kann man auch ihre Frequenz beobachten und vergleichen.

Fragestellung: *Welche Wörter haben die gemeinsame Wortwurzel* -grund-*?*

Die Recherche kann in jedem Korpus durchgeführt werden, solange die Suchanfrage nach einem Intervall (Zeichenkette) möglich ist. Hier werden zwei Recherchen durchgeführt: im **InterCorp**, denn hier können auch Angaben über die vertretenen Wortarten abgerufen werden, und dann im **DeReKo**, das aufgrund seiner Größe wesentlich mehr Formen liefern kann.

Recherche im InterCorp

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query Type: Character auswählen
- Character (Suchfeldeingabe)**: grund** → Search

Alternative:

- Query Type: Word Form
- Word Form (Suchfeldeingabe): **.*grund.*** → Search

Nach dem die Konkordanzen erschienen sind:

- Frequency → Node forms (Ergebnisse siehe Abb. 102)

dann

- Frequency → Node tags (Ergebnisse siehe Abb. 103)

Die Ergebnisse erscheinen in Form einer Liste, gereiht nach relativer Frequenz der KWIC-Formen (Abb. 102) oder nach der morphosyntaktischen Charakteristik der KWIC (Tag) (Abb. 103).

Frequenzliste der Formen (gekürzt)

	Filter	Tag	Abs.	%
1.	p/ n	aufgrund	16,283	35.9
2.	p/ n	zugrunde	3,642	8.0
3.	p/ n	grundlegend(e)	3,400	7.1
4.	p/ n	Aufgrund	2,814	6.2
5.	p/ n	Hintergrund	2,797	6.2
6.	p/ n	Rechtsgrundlage	2,330	5.1
7.	p/ n	grundsätzlich	1,824	4.0
8.	p/ n	Rechtsgrundlagen	786	1.7
9.	p/ n	Erwägungsgrund	714	1.6
10.	p/ n	Vordergrund	580	1.3

Abb. 102: Abfrage *-grund-*, Node forms (alle Texte)

Frequenzliste der Tags

	Filter	Tag	Abs.	%
1.	p/ n	APPR	19,097	42.1
2.	p/ n	NN	12,524	27.6
3.	p/ n	ADJA	7,323	16.1
4.	p/ n	ADV	3,338	7.4
5.	p/ n	ADJD	2,695	5.9
6.	p/ n	PTKVZ	298	0.7

Abb. 103: Abfrage *-grund-*, Node tags

Aus den Kürzeln der Tags in der Abb. 103 (Erklärung siehe im Kap. 8.2) ist ersichtlich, dass die Wortwurzel *-grund-* überraschenderweise am häufigsten in der Präposition *aufgrund* vorkommt. Erst danach haben in diesem Korpus die zweit höchste Frequenz Substantive (Komposita *Hintergrund*, *Rechtsgrundlage*, *Erwägungsgrund* und *Vordergrund*), Adjektive (konvertiertes Partizipium *grundlegend*, Ableitung vom Substantiv *grundsätzlich*). Man muss jedoch immer im Auge behalten, in welchen Texten recherchiert wurde: Hier waren es verschiedene Texte der geschriebenen Sprache, noch dazu bilden den größten Teil Gesetzestexte der EU und Schriften aus dem EU-Parlament, publizistische Texte und andere Sachtexte.

In vorwiegend belletristischen Texten (InterCorp: Kernkorpus) sieht die Reihenfolge etwas anders aus. Die nachfolgende Recherche deutet darauf hin, dass die Präposition *aufgrund* eigentlich typisch für Sachtexte ist und in belletristischen Texten nicht so häufig vorkommt. Immerhin nähern sich belletristische Texte der gesprochenen Sprache stärker an als Sachtexte, beispielswiese in der direkten Rede. Dies bestätigte auch eine Recherche in der DGD (Korpus nur mit gesprochener Sprache): Die Präposition *aufgrund* kommt etwa 10 Mal seltener vor als das konkurrierende *wegen*.

Frequenzliste der Formen (gekürzt)

	Filter	Tag	Abs.	%
1.	p/ n	Hintergrund	668	20.5
2.	p/ n	zugrunde	324	10.0
3.	p/ n	Abgrund	309	9.5
4.	p/ n	aufgrund	280	8.6
5.	p/ n	grundsätzlich	192	5.9
6.	p/ n	grundlegende	124	3.8
7.	p/ n	grundlegenden	117	3.6
8.	p/ n	Vordergrund	107	3.3
9.	p/ n	Untergrund	88	2.7
10.	p/ n	grundlegend	70	2.2

Abb. 104: Abfrage *-grund-*, Node forms (Belletristik)

Frequennzliste der Tags

	Filter	Tag	Abs.	%
1.	p/ n	NN	1,697	52.2
2.	p/ n	ADJA	511	15.7
3.	p/ n	ADJD	378	11.6
4.	p/ n	APPR	327	10.1
5.	p/ n	ADV	271	8.3
6.	p/ n	PTKVZ	51	1.6

Abb. 105: Abfrage *grund*, Node tags

Wenn man die Abb. 102 und 104 vergleicht, sieht man die Verschiebung der Reihenfolge in der Frequenz einzelner Substantive.

Diese Ergebnisse können an Texten des größten Korpus des Deutschen **DeReKo** überprüft werden.

Recherche im DeReKo

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: **Suchmodalitäten:**
  - Groß- / Kleinschreibung beachten für 1. Zeichen **auswählen**
  - Groß- / Kleinschreibung beachten für andere Zeichen abwählen
  - **Expansionslisten anzeigen:** **auswählen**
  - **Expansionslisten**
    - **mit Häufigkeiten** (langsam) **auswählen**
  - Sortierung nach Häufig. **absteigend** **auswählen**

Zusammenfassung ab 100 **auswählen** → Übernehmen

Lemmatisierung:

Komposita	abwählen
Sonstige Wortbildungsformen	**auswählen**
Spezialfälle	abwählen → Übernehmen

- **Suchanfrage** (Eingabe ins Suchfeld)**: &Grund** → Suchen

Die Abfrage bedeutet: Suche alle Formen zur Grundform *Grund*.
In der Liste der Ergebnisse (Abb. 106) ist auch das Basissubstantiv *Grund* aufgelistet (im Unterschied zum InterCorp, wo dieses ausgeblendet wird). Die Wörter in den Listen in Abb. 102, 104 und 106 sind fast gleich (Differenzen sind auf geringe Frequenzunterschiede zurückzuführen).

Formenliste zur Abfrage Lemma: *Grund* (gekürzt)

```
Suchbegriff-Expansionslisten (nach Häufig. absteigend)
Grund                : 877.086 (41.20%)
Hintergrund          : 318.064 (14.94%)
Gründen              : 295.670 (13.89%)
Gründe               : 199.128 (9.35%)
Aufgrund             : 170.052 (7.99%)
Grunde               :  94.000 (4.42%)
Hintergründe         :  39.966 (1.88%)
Untergrund           :  33.907 (1.59%)
Hauptgrund           :  19.510 (0.92%)
Abgrund              :  14.356 (0.67%)
HINTERGRUND          :   9.324 (0.44%)
Grundig              :   9.188 (0.43%)
Abgründe             :   7.076 (0.33%)
grundlos             :   6.939 (0.33%)
Hauptgründe          :   5.422 (0.25%)
```

Abb. 106: Formenliste (Expansionsliste) der Wortfamilie *-grund-* (DeReKo)

In die Expansionsliste (Abb. 106) ist auch ein „blinder Passagier" mittransportiert worden: die Marke *Grundig*. An diesem Fehler sieht man, wie das Korpus funktioniert: es sucht nach entsprechenden Graphemketten und vergleicht sie mit einem Wörterbuch bzw. einer Liste der Lemmata, über die das Korpus lemmatisiert wurde. In diesem Wörterbuch fehlte offensichtlich der Name dieser Firma.

Auch DeReKo ermöglicht die Abfrage nach einem offenen Intervall, also nach Wortformen, in denen die Buchstabenkombination *grund* vorkommt:
Einstellung wie vorher

- **Suchanfrage** (Eingabe ins Suchfeld)**: *grund*** → Suchen

Diese Abfrage bedeutet: Suche nach allen Wörtern, die die Buchstabenkombination *grund* beinhalten. Die Errechnung der Ergebnisse dauert im DeReKo sehr lange. Im Prinzip kann nur die Expansionsliste (wie Abb. 106) abgerufen werden – sie zeigt mehr als 68.000 Formen. Ihr Inhalt entspricht den Ergebnissen in der Abb. 104 (nach Frequenz absteigend: *aufgrund, Hintergrund, gegründet, grundsätzlich, Vordergrund, gründlich* etc.)
Es ist nicht sinnvoll, die Ergebnisse (Konkordanzen) abzurufen, nachdem die Expansionsliste erschienen ist, weil die Dauer der Erstellung der gesamten KWIC-Liste das erlaubte Limit

überschreitet, der Korpusmanager COSMAS II liefert eine Fehlermeldung. Darüber hinaus kann man mit Konkordanzen, in denen gemischt alle Wörter einer Wortfamilie vorkommen, wenig anfangen. So gesehen erweist sich die Arbeit mit InterCorp als effektiver.

Fazit:
Vergleicht man die Ergebnisse der Recherche im InterCorp mit DeReKo, stellt man fest, dass neben dem Stamm *Grund* (mit allen seinen Deklinationsformen) seine folgenden Ableitungen sehr häufig (daher für die Lernenden auch wichtig) sind: *aufgrund*; *grundlegend(e/en/er)*; *grundsätzlich*; *Hintergrund*; *zugrunde*. Beim letzten Wort handelt es sich um einen Verbzusatz zu *gehen*/ *liegen*/ *richten*. Verbzusätze werden in der Studie 7 behandelt.

Schlussbemerkung zur Studie 5

Für die Ermittlung der Wortformen ist InterCorp bisher die beste Wahl. Ohne komplizierte Einstellungen kann man alle Formen abrufen und dann auf einen Klick ihre Frequenzen berechnen lassen. Der Vorteil liegt auch darin, dass die Frequenzliste sich durch Klicken auf word in eine alphabetische Liste umwandeln lässt.

Die zentralen Glieder der Wortfamilie *grund* (zwanzig mit der höchsten Frequenz) sind:

> *abgrund, abgrundtief, aufgrund, beweggrund, großgrundbesitz, großgrundbesitzer, grundlegend, grundlos, grundsätzlich, grundverschieden, hauptgrund, hintergrund, lebensgrundlage, meeresgrund, untergrund, untergrundbahn, urgrund, vordergrund, zugrunde, zugrundeliegend*

Groß-/Kleinschreibung wurde hier ignoriert. Dies könnte jetzt im Unterricht ein guter Anlass sein, in einer Grammatikarbeit die Lernenden entscheiden zu lassen, zu welcher Wortart die einzelnen Wörter gehören.

Im akademischen Bereich können an solchen Wortfamilien die einzelnen (wichtigsten) Wortbildungsarten des Deutschen gezeigt werden:

Komposition: *Grund* → als Bestimmung *Grundbesitz, grundlegend*
→ als Basis *Beweggrund, Meeresgrund*; (weiter *Hauptgrund*, *Hintergrund*, *Untergrund*, *Vordergrund* als Übergangsbereich zur Derivation)

Derivation: *Grund* → *Abgrund, Urgrund, grundlos*

Konversion: *auf Grund*, *zu Grunde* (Präpositionalverbindungen) → *aufgrund* (Präp.); *zugrunde* (Part. (Verbzusatz)); *Grund* (Subst.) → *gründen* (Verb)

Zu jedem Wort können mehrere Sätze gezeigt werden, die die Bedeutungsunterschiede beleuchten.

Studie 6: Wortbildung

Die Wortbildung stellt einen Bereich der Sprache dar, der mithilfe von Computerprogrammen bestens erforscht werden kann. Jeder Texteditor und fast jedes Internetformular oder jede Suchmaschine kann nach genauen Wörtern oder nach Wortteilen suchen[38]. Versteht man die Wortbildung als „Bildung eines Lexems bis zur Wortgrenze", dann bieten Korpora unzählige Möglichkeiten an, Komposita und Ableitungsformen zu untersuchen. Mit einer gut formulierten Abfrage kann man sehr schnell einen Überblick bekommen, wie sich Wortstämme, Vor- und Nachsilben (Affixe) verhalten, welche Wörter zusammengesetzt werden können, welche Wortkombinationen üblich sind.

Erfahrungsgemäß ist für Deutschlerner/-innen oft die Wahl des richtigen adjektivischen Suffixes problematisch: „Wie heißt es eigentlich: *lustig, lustich, lustlich* oder gar *lustisch*?" Als ein Fallbeispiel der Wortbildung werden die Adjektivableitungen untersucht.

Fragestellung: *Welche Grundwörter verbinden sich mit den adjektivischen Suffixen* -ig, -lich *und* -isch*?*

Die Suche kann in jedem Korpus erfolgen. Wahrscheinlich am schnellsten und einfachsten geht sie im **InterCorp**.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query Type: CQL auswählen
- CQL (Suchfeldeingabe): **[tag="ADJ.*"&lemma=".*[ig|lich|isch]"]** → Search

Die Abfrage bedeutet: „Suche alle Adjektive, die auf *-ig, -lich* oder *-isch* enden". Sie liefert fast 2 Mio. Konkordanzen, in denen viele KWICs gleich sind, weil die Wörter mit Ableitungen auf *-ig, -lich* oder *-isch* ziemlich häufig sind: *wirklich* (zu *wirken*), *plötzlich* (zu *Plotz*[39]), *völlig* (zu *voll*), *möglich* (zu *mögen*), um nur die häufigsten vier zu nennen. Man braucht jedoch jedes Wort, besser sogar nur seine Grundform (Lemma), nur einmal, um festzustellen, mit welchem Suffix (Nachsilbe) es vorkommt. Dies erreicht man über die Frequenzliste:

- Frequency → Custom → Level 1 → Attribute: Lemma auswählen → ignore case auswählen → Position: Node auswählen → Make Frequency List

Die Ergebnisse können in eine Excel-Tabelle übertragen und alphabetisch sortiert werden. In der alphabetischen Liste wurden alle Ableitungen mit gleichem Wortstamm markiert und diese in die Tab. 25 (auf der folgenden Seite) übertragen.

[38] Einige Suchmaschinen können auch nach Lemmata suchen (Google), manche sogar nach Allonymen (Eigennamen, die in verschiedenen Sprachen anders aussehen): Auf dem Portal der Österreichischen Bundesbahnen (www.oebb.at) kann man z.B. die Suche nach der Verbindung *Brünn Hbf. – Preßburg-Engerau* eingeben und bekommt automatisch den Fahrplan zwischen *Brno hl.n.* und *Bratislava-Petržalka*.

[39] Bedeutete „Aufprall" (Kluge 2002:709).

Adjektivische Ableitungen mit derselben Basis

Lemma	abs. Freq.
geistig	2413
geistlich	208
geschäftig	114
geschäftlich	690
gütig	209
gütlich	118
heimisch	505
heimlich	1170
herrisch	88
herrlich	1010

Lemma	abs. Freq.
herzigen[1)]	5
herzlich	1768
höfischen	21
höflich	892
jährlich	9101
-jährig[2)]	5450
kindisch	143
kindlich	358
launig	20
launisch	53

Lemma	abs. Freq.
mündig	48
mündlich	1615
ungläubig	440
unglaublich	1232
verständig	77
verständlich	1295
zeitig	140
zeitlich	1515

Tab. 25: Adjektivische Ableitungen (InterCorp)

Anmerkungen:
[1)]Signifikante Frequenz hat nur *halbherzig*, (*un*)*barmherzig*.
[2)]Kommt nur als Basis eines Kompositums vor: *mehrjährig*, *langjährig*, *diesjährig*, *dreijährig*, *minderjährig*, *fünfjährig* sind die häufigsten aus insgesamt 203 Komposita im InterCorp.

Die 28 Lexeme in der Tabelle 24 wurden aus 3.400 Lemmata (Grundformen) von Adjektiven auf *-ig*, *-lich* oder *-isch* aussortiert. Ihre Frequenz im Korpus InterCorp_de (Gesamtkorpus) ist (bis auf *herzig*) höher als 20 Belege in absoluten Zahlen.

Fazit:
Kein Grundwort bildet frequentierte Ableitungen mit allen drei Suffixen. Es gibt auch sehr wenige Paare (14), die zwei von diesen Suffixen haben, wobei beide suffigierten Wörter einigermaßen häufig vorkommen.

Aus der Tab. 24 geht auch hervor, dass jedes von diesen Suffixen das jeweilige Wort so weit spezifiziert, dass es zwischen zwei Ableitungen von derselben Basis gravierende semantische Unterschiede gibt: *geistig – geistlich; höflich – höfisch*; … *zeitig – zeitlich*. Der minimalste Unterschied besteht vielleicht noch zwischen *launig* und *launisch*.

Unter den aufgelisteten Ableitungen fehlt z.B. *neidig* x *neidisch*. *Neidisch* hat (im InterCorp) relativ niedrige Frequenz (nur 52 Belege in absoluten Zahlen), *neidig* kommt gar nicht vor. Im DeReKo ist das Verhältnis *neidig* zu *neidisch* etwa 1 zu 4. *Neidisch* ist also – allgemein gesehen - die üblichere Form in geschriebenem Deutsch. Die Länderansicht im DeReKo zeigt jedoch deutlich, dass die Form *neidig* eindeutig in österreichischen Texten überwiegt. Während *neidisch* in deutschen und schweizerischen Texten gleich vertreten ist, kommt diese Form in österreichischen Texten signifikant seltener vor.

Schlussbemerkung zur Studie 6:
Diese Unterschiede zwischen den einzelnen Paaren in dieser Studie könnte man in einem weiteren Rechercheschritt erkunden: über die Kollokationsanalyse bzw. über den Kontrast zu einer anderen Sprache. (Mehr zu Wortbildung im Kontrast in Káňa 2012).

Studie 7: Verben mit Zusatz

Diese Problematik liegt zwischen den Fachgebieten der Morphologie, Syntax und Lexikologie. Als Zusatz zum Verb wird hier der trennbare Teil des Verbalkompositums verstanden, der traditionell oft als „trennbares Präfix“ (in Opposition zum „untrennbaren Präfix“)[40] bezeichnet wird. Dieses Phänomen bereitet den Deutschlernenden ziemliche Schwierigkeiten und „sollte im DaF/DaZ-Unterricht dementsprechend berücksichtigt werden“ (Fachlexikon DaF/DaZ 2010: 344). Schwierig empfinden die Lernenden vor allem homographische Verben (mit gleichem Schriftbild, jedoch unterschiedlich betont): *umstellen* vs. *umstellen*, *übersetzen* vs. *übersetzen*, aber auch andere Verben, bei denen (oft auch Muttersprachler/-innen) zweifeln, ob der Zusatz abgetrennt werden soll oder nicht (z.B. (*sich*) *widerspiegeln*).

Die Verbzusätze lassen sich in Korpora leicht abfragen, wenn man das Tagging in morphosyntaktisch annotierten Korpora zur Hand hat. Da die meisten Korpora für Deutsch mit dem gleichen Tagset annotiert sind, reicht es, wenn man sich merkt, dass Verbzusätze das Kürzel PTKVZ haben.

Fragestellung 1: *Was sind die häufigsten Verbzusätze im Deutschen?*

Recherchiert wird in mehreren Korpora: **DWDS**, **DeReKo** und **InterCorp**.

Recherche im DWDS

- **DWDS** → Abfragefenster
- Suchfeldeingabe: **$p=PTKVZ** → oder Suche im DWDS

Das Ergebnis sind Konkordanzen der einzelnen Korpora des DWDS. In der Abb. 107 sind lediglich 4 Beispiele aus insgesamt 721.351 Treffern im Kernkorpus angeführt:

… man deutet **an**, daß der Nobelpreis einem aufgrund eines Irrtums verliehen wurde.
»[…] Wart's **ab**«, sagt die Schrift und fährt **fort**: » … in der Straßenbahn, die vollbesetzt war, … «
In der mündlichen Kommunikation dagegen kommt es nicht primär auf Sachlichkeit **an**, sondern auf …
Sie beschwören dann durch ein paar Zitate eine Vision der gerade erlebten Situation **hervor** …

Abb. 107: Konkordanzen zur Abfrage nach abgetrennten Verbzusätzen (DWDS)

Es ist im DWDS nicht möglich, die Treffer auf nur einen Beleg pro Form zu reduzieren, um festzustellen, welche Zusätze es gibt. Eine Statistik der Frequenzen einzelner Zusätze lässt sich im DWDS auch nicht erstellen. Aus diesen Gründen wird in weiteren Korpora recherchiert.

[40] Vgl. Engel 1988: 440 oder Zapletal et al. 1980: 128.

Recherche im DeReKo

Dieselbe Recherche kann man auch im **DeReKo** durchführen. Dazu wird **COSMAS II$_{win}$** dringend empfohlen, denn nur hier kann auch die Statistik der Häufigkeiten von Verbzusätzen abgerufen werden.

- **Cosmas II$_{win}$** → Recherche →
- Archiv: TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
- Korpus: TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T
- **Suchanfrage** (Eingabe ins Suchfeld): **MORPH(PTK vz)** → suchen

Die Abfrage liefert 6,5 Mio. Treffer, viele von ihnen haben dieselbe Form, und diese Form (Wort-Type) will man nur einmal ansehen. Dazu dient der Befehl:

- Ansicht → §Ansicht nach Wort-Types (nur im COSMAS II$_{win}$!)

Das Ergebnis – 405 Wort-Types - ist in der Abb. 108 zu sehen. Sie sind nach der relativen Frequenz (**rel [%]**) absteigend sortiert, können jedoch auch nach der absoluten Frequenz oder alphabetisch neu sortiert werden, wenn man die entsprechende Überschrift im Kopf der Tabelle anklickt.

Wort-Types zur Abfrage Verbzusätze (gekürzt)

```
Ergebnisübersicht
Ansicht nach Wort-Types, 405 Wort-Types, nach »rel [%]« absteigend sortiert.
Anz Treffer Anz Texte rel [%] Wort-Type (Eb+Rb+Db+Si)
801.659   671.297  12.198 an
678.273   572.282  10.321 aus
583.408   505.005   8.877 ein
490.089   430.096   7.457 ab
485.011   423.399   7.380 auf
398.957   320.681   6.070 statt
369.695   331.699   5.625 vor
300.784   266.270   4.577 zurück
234.199   214.855   3.564 zu
199.533   188.191   3.036 mit
153.481   143.023   2.335 fest
126.823   120.459   1.930 weiter
126.329   119.641   1.922 zusammen
107.382   101.748   1.634 hin
91.608    85.979   1.394 durch
81.494    78.666   1.240 nach
72.302    69.306   1.100 heraus
69.599    67.228   1.059 entgegen
63.334    60.796   0.964 her
57.526    54.570   0.875 teil
56.625    53.823   0.862 vorbei
54.986    53.187   0.837 bei
49.291    47.490   0.750 um
48.116    46.580   0.732 hervor
47.617    46.238   0.725 hinaus
43.285    40.990   0.659 dar
42.779    41.179   0.651 hinzu
39.550    38.162   0.602 da
36.923    35.514   0.562 los
33.349    31.934   0.507 weg
```

Abb. 108: Frequenz der Formen zur Abfrage: Verbzusätze (DeReKo, Archiv TAGGED)

Die Verben zu einzelnen Zusätzen findet man wie folgt: mit der Maus auf die Zeile mit dem gewünschten Zusatz fahren, mit einem Doppelklick werden die Konkordanzen abgerufen, in denen als KWIC der Verbzusatz dargestellt wird (in der Abb. 109 ist es *fort*). Nach dem Verb,

zu dem der Verbzusatz gehört, muss entweder in der Konkordanzzeile gesucht werden oder es kann über die Kookkurrenzanalyse ermittelt werden (siehe weiter).

A09 in Bleichi und setzte seine Fahrt	**fort,** ohne sich um den Schaden zu kümmern.
A09 dennoch nicht abwenden», fährt er	**fort.** Zu Hause in den warmen Winterstuben
A09 hnson setzt 1963 ein, schreitet	**fort** bis ins Jahr 1970 und endet mit einem
A09 Damit setzte sich der Strukturwandel (…)	**fort.** Seit 1999 hörten 402 Bauern
A09 hübsch und sehr gescheit», fährt Leonor	**fort.** Sie bricht in Tränen aus.

Abb. 109: Konkordanzen zum Verbzusatz *fort* (Auswahl aus 28.671 Treffern)

Mit nur kursorischem Lesen dieser Konkordanzzeilen fällt auf, dass das häufigste Verb mit dem Zusatz *fort* vermutlich *fortsetzen* sein wird, gefolgt von *fortfahren* und *fortschreiten*.

Recherche im InterCorp

Noch einmal werden die Verbzusätze im **InterCorp** abgerufen.

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query type: CQL
- CQL (Suchfenstereingabe)**: [tag="PTKVZ"]** → Search

Nachdem die Konkordanzen erschienen sind:

- Frequency → Node forms

Als Ergebnis bekommt man eine Frequenzliste der Verbzusätze nach Frequenz absteigend (gekürzt):

		word	Freq	Freq [%]
1.	p/ n	an	38,280	10.8
2.	p/ n	auf	34,032	9.6
3.	p/ n	aus	27,689	7.8
4.	p/ n	vor	24,259	6.8
5.	p/ n	zu	20,923	5.9
6.	p/ n	ein	19,950	5.6
7.	p/ n	ab	18,122	5.1
8.	p/ n	zurück	14,344	4.0
9.	p/ n	mit	11,035	3.1
10.	p/ n	fest	9,764	2.7
11.	p/ n	hin	9,270	2.6
12.	p/ n	zusammen	6,715	1.9
13.	p/ n	nach	6,215	1.7
14.	p/ n	dar	5,934	1.7
15.	p/ n	um	5,532	1.6

		word	Freq	Freq [%]
16.	p/ n	weiter	5,140	1.4
17.	p/ n	heraus	4,071	1.1
18.	p/ n	fort	3,846	1.1
19.	p/ n	hervor	3,808	1.1
20.	p/ n	hinaus	3,667	1.0
21.	p/ n	da	3,575	1.0
22.	p/ n	her	3,559	1.0
23.	p/ n	durch	3,485	1.0
24.	p/ n	herum	3,446	1.0
25.	p/ n	bei	3,305	0.9
26.	p/ n	statt	3,285	0.9
27.	p/ n	entgegen	2,507	0.7
28.	p/ n	hinzu	2,466	0.7
29.	p/ n	vorbei	2,321	0.7
30.	p/ n	hinein	2,124	0.6

Abb. 110: Frequenzliste der Verbzusätze (InterCorp)

Die einzelnen Posten im Kopf der Tabelle kann man anklicken und so nach diesem Argument neu sortieren. Nach dem Klicken auf **word** werden die Treffer alphabetisch sortiert.

Fazit:
Die Rechercheergebnisse aus DeReKo und InterCorp sind durchaus vergleichbar: In den Abb. 108 und 110 sind die 30 häufigsten Verbzusätze im jeweiligen Korpus dargestellt. Die Verbzusätze *fort-*, *herum-*, *hinein-*, *los-*, *teil-*, und *weg-* fehlen zwar in einer der beiden Ergebnislisten. Sie kommen jedoch jeweils in den unmittelbar nachkommenden Positionen (31 bis 60) vor. Vergleicht man also die häufigsten 60 Verbzusätze in beiden Korpora, ergibt sich das Bild über die häufigsten Verbzusätze in geschriebenem Deutsch:
ab-, an-, auf-, aus-, bei-, da-, dar-, durch-, ein-, entgegen-, fest-, fort-, her-, heraus-, herum-, hervor-, hin-, hinaus-, hinein-, hinzu-, los-, mit-, nach-, statt-, teil-, um-, vor-, vorbei-, weg-, weiter-, zu-, zurück- und *zusammen-*.

In einem weiteren Schritt wird festgestellt, welche Verben diese häufigsten Verbzusätze haben.

Fragestellung 2: *Zu welchen Grundverben tritt der häufigste Verbzusatz (*an*) zu?*

Der Anfang der Recherche hat denselben Ablauf wie in der vorherigen Fragestellung. Die Fortsetzung wird jedoch nur im **InterCorp** durchgeführt.

Nachdem die Liste mit den häufigsten Verbzusätzen abgerufen worden ist (Abb. 110), klickt man in der zweiten Spalte auf p (= positiver Filter), wie der Ausschnitt aus der Liste (Abb. 110a) zeigt:

		word	Freq	Freq [%]
1.	**p**/ n	an	38,280	10.8
2.	**p**/ n	auf	34,032	9.6
3.	**p**/ n	aus	27,689	7.8
4.	**p**/ n	vor	24,259	6.8

Abb. 110a: Ausschnitt aus der Frequenzliste der Verbzusätze (Abb. 110, vorherige Seite)

Danach erscheinen unsortierte Konkordanzen, in denen **an** als KWIC steht (am Ende des Satzes). Die Verben zum Verbzusatz *an* lassen sich am effektivsten über die Kollokationsberechnung ermitteln:

- Collocation → Custom…

Einstellung der Kollokationsanalyse (Collocations candidates):

- Attribute: Lemma
- In the range from: -10 to 0
- Minimum frequency in corpus: 10 (spielt keine Rolle für diese Abfrage)
- Minimum frequency in given range: 10 (spielt keine Rolle für diese Abfrage)

Make Candidate List

Aus der Liste der insgesamt 5.396 automatisch berechneten Kollokatoren (in der Abb. 111 auf der folgenden Seite siehe links oben die Angabe Total) müssen alle Nicht-Verben manuell aussortiert werden (in der Abb. 111 sind sie durchgestrichen).

Kollokationspartner zum Verbzusatz *an* (gekürzt)

Total: 5396											
			Freq	**log likelihood**	**logDice**				**Freq**	**log likelihood**	**logDice**
1.	p/n	**sehen**	5249	42336.240	10.649	16.	p/n	**ziehen**	735	5007.362	8.773
2.	p/n	**fangen**	2028	25391.592	10.628	17.	p/n	**halten**	930	5589.093	8.766
3.	p/n	**nehmen**	2637	20227.721	10.100	18.	p/n	**hören**	673	4390.259	8.622
4.	p/n	**starren**	1261	14800.071	9.966	19.	~~p/n~~	~~**sie**~~	~~4343~~	~~19057.474~~	~~8.503~~
5.	p/n	**schauen**	1001	9907.180	9.557	20.	~~p/n~~	~~**ich**~~	~~6214~~	~~27162.864~~	~~8.500~~
6.	p/n	**blicken**	954	9084.744	9.465	21.	p/n	**kündigen**	431	5003.835	8.487
7.	p/n	**rufen**	1005	8173.972	9.353	22.	~~p/n~~	~~**sich**~~	~~4075~~	~~17745.623~~	~~8.476~~
8.	p/n	**erkennen**	922	7679.240	9.288	23.	p/n	**zünden**	419	5932.286	8.469
9.	p/n	**wenden**	841	7341.082	9.231	24.	~~p/n~~	~~**Auge**~~	~~645~~	~~3900.646~~	~~8.465~~
10.	p/n	**bieten**	825	6780.409	9.145	25.	~~p/n~~	~~**darauf**~~	~~653~~	~~3664.029~~	~~8.353~~
11.	~~p/n~~	~~**er**~~	~~7774~~	~~40474.744~~	~~9.066~~	26.	~~p/n~~	~~**du**~~	~~1274~~	~~5797.522~~	~~8.327~~
12.	~~p/n~~	~~**«**~~	~~2839~~	~~15050.463~~	~~9.011~~	27.	~~p/n~~	**?**	1640	7098.896	8.293
13.	p/n	**kommen**	1600	8943.576	8.946	28.	p/n	**lächeln**	406	3423.408	8.291
14.	~~p/n~~	~~**Sie|sie**~~	~~3123~~	~~15952.103~~	8.920	29.	~~p/n~~	~~**“**~~	~~975~~	~~4517.483~~	~~8.245~~
15.	p/n	**schließen**	718	5279.247	8.854	30.	~~p/n~~	~~**wieder**~~	~~785~~	~~3858.788~~	~~8.244~~

Abb. 111: Frequenzliste der Kollokatoren zum Verbzusatz *an* (InterCorp)

Als Beispiel wird das zweit häufigste Verb in der Liste abgerufen: *anfangen*. Durch Klicken auf p (=positiver Filter) neben einem Verb (fangen) kann man Konkordanzzeilen mit dem Verb *anfangen* mit abgetrenntem *an* abrufen:

… fängt er endlich	**an**	, die Notwendigkeit (…) zu betonen .
… , sie ließ sich ihre Körbe bringen und fing	**an**	wie eine Gemüsefrau auf dem Markt …
Dies fängt bereits bei der Wortwahl (…)	**an**	, wobei bekanntlich schon die …
… , zeigte in die Ferne und fing	**an**	, mir (…) etwas zu erklären .
Dann fing ich	**an**	, es wirklich zu genießen
Wir fangen immer damit	**an**	, öffentliche Ausgaben zu beschneiden
, und sofort fing sie	**an**	zu dichten :

Abb. 112: Auswahl aus Konkordanzen zum positiven Filter des Kollokators *fangen* in der Abb. 111 (InterCorp)

Fazit:
Der Verbzusatz an kommt am häufigsten mit diesen Verben vor: *sehen* (*ansehen*), *fangen* (*anfangen*), *nehmen* (*annehmen*), *starren* (*anstarren*), *schauen* (*anschauen*), *blicken* (*anblicken*), *rufen* (*anrufen*), *erkennen* (*anerkennen*), *wenden* (*anwenden*), *bieten* (*anbieten*) und *kommen* (*ankommen*).

Analog können auch Verben mit anderen Zusätzen festgestellt werden. (Verben, zu denen die zehn häufigsten Verbzusätze zutreten, sind in der im Kap. 7.2 aufgelistet.) Zu allen können auch parallele Passagen in einer anderen Sprache angezeigt werden, falls dies am Anfang der Recherche eingestellt wurde, wie die folgende Fragestellung 3 zeigt.

Fragestellung 3: *Was sind die häufigsten Verben mit dem Zusatz* fest*? Was entspricht dem deutschen Verb* festhalten *im Englischen?*

Recherchiert wird im **InterCorp**, es gibt mehrere Wege, wie man das ganze Konjugationsparadigma des Verbs *festhalten* (analog aller anderen Verben mit trennbarem Zusatz) abrufen kann:

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Aligned corpora → intercorp_en → Add
  Query type: CQL
- CQL (Suchfenstereingabe): **[tag="PTKVZ"&word="fest"]** → Search

Nachdem die Konkordanzen erschienen sind:

- Collocations → Custom…

Einstellung der Kollokationsanalyse (Collocations candidates):

- Attribute: Lemma
- Range: -10 to 0 → Make Candidate List

Das Ergebnis dieser Abfrage ist eine automatisch erstellte Liste, in der auch andere Kollokationspartner erscheinen - Wörter, die signifikant oft mit dem Verbzusatz *fest-* vorkommen. Diese müssen ausgeblendet werden, denn für die Auswahl sind nur Verben relevant. Die manuell ausgewählten Verben sind in der Abb. 113 aufgelistet und nach der Wahrscheinlichkeit des Vorkommens mit *fest* (Signifikanzmaß logDice) sortiert.

			Freq	logDice				Freq	logDice
1.	p/n	stellen	4043	11.305	12.	p/n	klammern	48	7.264
2.	p/n	legen	1528	10.727	…				
3.	p/n	halten	921	9.506	29.	p/n	binden	38	6.696
4.	p/n	setzen	319	8.309	…				
5.	p/n	stehen	543	8.230	36.	p/n	sitzen	64	6.584

Abb. 113: Basisverben mit dem Verbzusatz *fest* (Auswahl)

Aus dieser Tabelle lassen sich die häufigsten Verben mit dem Zusatz *fest* ableiten: *feststellen, festlegen, festhalten, festsetzen, feststehen, festklammern, festbinden, festsitzen*.

Es muss betont werden, dass es sich in den Kollokationstabellen (auch Abb. 113) um potentielle Kandidaten handelt, die automatisch berechnet worden sind. Aus diesem Grund sollten immer die Konkordanzzeilen stichpunktartig durchgesehen werden, ob die Ergebnisse den Erwartungen entsprechen. Die folgenden Konkordanzzeilen zeigen, welche Ergebnisse zu erwarten sind, wenn man sich auf das Verb *sitzen* (Abb. 113) konzentriert und die Konkordanzen durch Klicken auf p abruft:

(1) wir **saßen** in der sowjetisch besetzten Zone **fest** , Mutter sogar freiwillig ,
(2) Im Augenblick **sitzen** Sie hier **fest** , weil wir Sie nicht auf Kaution
(3) eine Tatsache " , sagte Kaninchen . " Du **sitzt fest** . "
(4) Er **saß** oft auf seiner Veranda und stellte **fest**, ob die Züge pünktlich vorbeifuhren.

Die meisten Konkordanzen beinhalten tatsächlich das verbale Kompositum *festhalten* (1), (2), (3). Es können aber auch „Fehler“ vorkommen, wie das Beispiel (4) zeigt. Dennoch sind die Ergebnisse überzeugend, die absolut überwiegende Menge ist richtig berechnet, man kann sie als tragfähig für eine weitere Forschung betrachten.

Fazit:
Die Verben *feststellen, festlegen, festhalten, festsetzen* und *feststehen* sind die häufigsten mit dem Zusatz *fest-*. Ein wenig geringer, wenn immer noch hochfrequentiert sind *festklammern, festbinden*, und *festsitzen*. Vergleichsweise seltener sind die Verben *feststecken, festmachen, festnageln, festkrallen, festbeißen, festdrücken* und *festsaugen* (in der Verbindung *sich an jmdn. festsaugen*). Andere Verben mit *fest-* haben verhältnismäßig geringe Frequenz in der geschriebenen Sprache.

Konkordanzzeilen werden wieder durch Anklicken des positiven Filters (p) abgerufen, so auch die Konkordanzen, in denen *festhalten* eine Klammer (*...hält ... fest.*) bildet.

Jetzt wird nach den englischen Entsprechungen zum Verb *festhalten* gesucht. Zuerst muss das ganze Konjugationsparadigma des Verbs abgerufen werden. Dies muss wegen der Trennbarkeit des Zusatzes *fest* in zwei Schritten erfolgen:

Schritt 1:
- Query type: CQL
- CQL (Suchfenstereingabe): **[lemma="halten"] []* [word="fest"] within <s id=".*"/>** → Search

Die Abfrage bedeutet: Suche alle Formen von *halten*, davon rechts im beliebigen Abstand das Wort *fest*, allerdings nur bis zur Grenze des Satzes. (Eine Liste der Abfragemöglichkeiten im Modus CQL ist im Kap.VII angeführt.) Diese Abfrage ist eine Erweiterung der vorherigen Abfrage um das Argument within (Suche innerhalb eines definierten Korpusabschnittes, in diesem Fall Satz).

Schritt 2:
- Query type: CQL
- CQL (Suchfenstereingabe): **[lemma="festhalten"]** → Search

Diesen zweiten Schritt kann man als Standard bezeichnen. So werden üblicherweise alle Formen einer Grundform abgerufen (vgl. dazu auch Studie 4). Als Ergebnisse bekommt man Konkordanzen mit allen Formen von *festhalten*, die als eine ununterbrochene Zeichenkette (ein Wort) im Korpus vorkommen. Ein Ausschnitt davon ist in der Abb. 115 zu sehen.

Die Konkordanzzeilen zu jeder Abfrage erscheinen gleich mit einer Parallelpassage im Englischen (Abb. 114 und 115 auf der nächsten Seite). Diese Sprache wurde eben am Anfang der Recherche ausgewählt (siehe vorherige Seite oben).

Parallelpassagen Deutsch – Englisch

	intercorp_de	intercorp_en
(1)	Sie drückte ihn an sich und **hielt** ihn eine lange Minute ganz **fest** .	She squeezed him , and held him tight .
(2)	Auch wenn mir das nicht gelingt , **halte** ich an meiner eigenen Meinung und meinem Urteil **fest** .	And if that doesn't work , I 'll have to stick with my own opinions and judgment .
(3)	Der Blick seiner blauen Augen **hielt** sie **fest** , während er allmählich wieder zu Atem kam .	Langdon caught his breath as his blue eyes held her firmly .
(4)	Sie **hielt** sich einfach nur an ihm **fest** , und das war bei Gott ein fantastisches Gefühl .	All she did was hang on , and , God , it felt great .
(5)	Trotzdem **halte** ich an ihnen **fest** , trotz allem , weil ich noch immer an das innere Gute im Menschen glaube .	Yet I cling to them because I still believe , in spite of everything , that people are truly good at heart .
(6)	Ich **hielt** sie **fest** ; hätte ich das nicht getan , dann wäre sie von der Schaukel gefallen .	I held her firmly ; if not , she would 've fallen onto the porch .
(7)	Sie verschränkte die Arme vor der Brust und **hielt** sich an den Ellbogen **fest** .	She crossed her arms over her middle and hugged her elbows .
(8)	Isebel war eine Konservative und **hielt** am alten Glauben **fest** .	Jezebel was a conservative , sticking to the old beliefs against the new ones .
(9)	Arthur rappelte sich hoch und **hielt** sich ängstlich **fest** .	Arthur struggled to his feet and hugged himself apprehensively .

Abb. 114: Auswahl aus den Konkordanzzeilen zur Abfrage *festhalten* (abgetrennt) und ihre Entsprechungen im Englischen (InterCorp_de-en)

	intercorp_de	intercorp_en
(10)	… das schon lange und für immer in Rubens' Erinnerung **festgehalten** war:	…, as drawn long ago in Rubens' memory:
(11)	, dessen Natur (die es determiniert, es **festhält** und seit der Tiefe der Zeiten durchdringt) …	whose nature (that which determines it, contains it, and has traversed it from the beginning of time) …
(12)	"Wir brauchen ihn nur **festzuhalten**", sagte Frodo."	'All we need is something to keep a hold on him,' said Frodo.
(13)	»Wir können ihn nicht ewig **festhalten**, ohne Anklage zu erheben, …	"We can't hold him indefinitely without arraigning him, …
(14)	Mit einem Ruck riß er sich von dem Mann los, der ihn am Ärmel **festhielt**.	He tore his arm away from the man, who was now holding on to his sleeve.
(15)	Vielleicht ist es besser, die Identität der vorhandenen oder gezeigten Personen **festzuhalten**, um nicht unendlich in diese schwimmenden Bezeichnungen verwickelt zu werden.	Perhaps it would be better, once and for all, to determine the identities of all the figures presented or indicated here , so as to avoid …

Abb. 115: Auswahl aus Konkordanzzeilen zur Abfrage *festhalten* (lemmatisiert) und ihre Entsprechungen im Englischen (InterCorp_de-en)

Fazit:

Aus den Parallelpassagen in den Abb. 114 und 115 ist ersichtlich, wie sich das deutsche Wort *festhalten* verhält. Im Kontext ist es ein Bestandteil der folgenden Phraseme und Chunks:

sich festhalten (9),

(*sich*) *an jmdm./ etw. festhalten* (2, 4, 5, 8, 14),
jmdn. /sich/ etw. an etw. festhalten (7),
jmdn. /etw. festhalten (1, 3, 6, 11, 12, 13, 15)
in … festgehalten sein (10).

Die englischen Entsprechungen decken interessante Facetten im sprachlichen Kontrast auf:

sich festhalten (9),	*hug oneself*
sich festhalten (9),	*hug oneself*
hromadný e-mail: kvíz, provoz, neplatění pokuty za 15.9.2014 (*sich*) *an jmdm./ etw. festhalten* (2, 4, 5, 8, 14)	*stick to* (2, 8), *hang on* (4), *cling to* (5), *hold on to* (14)
jmdn. /sich/ etw.an etw. festhalten (7),	*hug st.*
jmdn. /etw. festhalten (1, 3, 6, 11, 12, 13, 15)	*hold sb./st. tight* (1, 13) *firmly* (3, 6), *contain st.* (11), *to keep a hold on sb.* (12), *determine* (15)
in … festgehalten sein (10).	*drawn* (metaphorisch)

Für lexikographische Zwecke müsste jedes intersprachliche Paar noch gründlich untersucht und statistisch ausgewertet werden. Dennoch umreißen auch diese wenigen Pendants im Englischen ein Bild über die Entsprechungen, die ein herkömmliches Wörterbuch kaum liefern könnte. Diesen Vorteil nutzen eher erfahrene Lehrer/-innen, Linguist/-innen oder Student/-innen. Anfänger/-innen könnten mit diesem Überangebot an unterschiedlichen Entsprechungen natürlich überfordert sein.

Studie 8: Verbalkomplexe

Die Problematik der deutschen Verbalkomplexe wird in jeder Grammatik behandelt, in Lehrbüchern wird ihr auch relativ viel Platz gewidmet. Selten wird jedoch auf (regionale) Unterschiede in der Reihenfolge innerhalb der Verbalkomplexe hingewiesen.
Das folgende Phänomen ist zwar allgemein bekannt, in gängigen Grammatiken wird es allerdings nicht „plurizentrisch" behandelt und (meines Wissens) wurde es an einer repräsentativen Anzahl von Beispielen bisher noch nicht belegt. Das folgende Beispiel für die Korpusarbeit behandelt dreiteilige Verbalkomplexe am Ende eines Nebensatzes.
Nach Duden - Grammatik (2005: 480-482), Zifonun (1997: 1285-1287) oder Engel (1988: 445-447) ist die „grammatikalische" Reihenfolge eines dreiteiligen Verbalkomplexes wie folgt:
finites Auxiliar – Infinitiv Vollverb – Infinitiv des Modal-/oder AcI-Verbs[41], also prinzipiell wie in (1).

(1) *Das hätte bedeutet, dass die Bauarbeiten spätestens 2009* ***hätten beginnen müssen***.[42]

Andere Reihenfolgen, auch die in (2), wurden von Lehrern/-innen oft als „falsch" gekennzeichnet. Die österreichische sprachliche Realität kennt aber eben diese Abfolge:
Infinitiv Vollverb – finites Auxiliar – Infinitiv Modalverb:

(2) *»Nur das Geschwätz, das er halt* ***schreiben hat müssen****, sagt der Doktor Forster«, erklärte Valerie nervös…* [43]

Ob diese (2) Reihenfolge in Österreich zur Standard- oder Randerscheinung gehört, kann mit Korpusdaten belegt werden. Für die Recherche eignet sich besser das **DeReKo** als andere Korpora, denn hier sind die Ursprungsländer der Texte sehr einfach abrufbar und die Ergebnisse äußerst übersichtlich.

Recherche im InterCorp

Zuerst wird die Reihenfolge wie in (1) abgerufen (**Abfrage 1**), dann die Reihenfolge wie in (2) (**Abfrage 2**).

Abfrage 1: finites Auxiliar – Infinitiv Vollverb – Infinitiv des Modal-/oder AcI-Verbs

- **Cosmas II$_{web}$** → Recherche
- Archiv: TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
- Korpus: TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T
- **Optionen:**
  Expansionslisten abwählen → Übernehmen
  Ergebnispräsentation
  Länderansicht auswählen → Übernehmen
- **Suchanfrage** (Eingabe ins Suchfeld):
  MORPH-Assistent → Verben: finit, ohne Imperativ → Klasse: auxiliar auswählen → Übernehmen

[41] AcI = lat. *Accusativ cum Infinitivo*, grundsätzlich handelt es sich im Deutschen um Verben mit Infinitiv ohne *zu* (Wahrnehmungsverben, *lassen* etc.).
[42] St. Galler Tagblatt, 3.1.2009, S. 31 (DeReKo).
[43] Simmel, Johannes Mario: Und Jimmy ging zum Regenbogen (InterCorp).

Im Suchfeld erscheint: **MORPH(VRB fin a)**, dann den Cursor hinter die Klammer positionieren und mithilfe des morphologischen Assistenten das nächste Verb definieren:

MORPH-Assistent → Verben: Infinitiv → Klasse: voll auswählen → Übernehmen

Dieser Schritt ergibt im Suchfeld: **MORPH(VRB fin a)MORPH(VRB inf v)**. Jetzt den Cursor wieder hinter der Klammer positionieren und das letzte Verb definieren:

MORPH-Assistent → Verben: Infinitiv → Klasse: modal auswählen → Übernehmen

Nach diesem Schritt sind alle Verben im Suchfeld definiert, die Abfrage sieht folgend aus:

MORPH(VRB fin a)MORPH(VRB inf v)MORPH(VRB inf m)

Nun fehlt die rechte Satzgrenze zu markieren, wo sich der Verbalkomplexe befinden sollen:

#IN(R) <s>

Dieser Schritt eliminiert aus den Konkordanzen zufällige Anhäufungen von Verben.

Die ganze Eingabe lautet jetzt (**Achtung!** Leerzeichen an den richtigen Stellen hinzufügen!):

MORPH(VRB·fin·a) MORPH(VRB·inf·v)·MORPH(VRB·inf·m) #IN(R) <s> → Suchen

Die Abfrage bedeutet: Suche finites Hilfsverb (VRB fin a), Infinitiv eines Vollverbs (VRB inf v) und Infinitiv eines Modalverbs (VRB inf m) in dieser Reihenfolge und am Ende (#IN(R) =„rechtes Ende“) eines Satzes (<s>).

Das Ergebnis sind über 20.000 Treffer, aus denen in die Abb. 116 per Zufall 10 Konkordanzzeilen ausgewählt wurden.

Verbalkomplexe im (DeReKo, Tagged-T)

```
KWIC (unsortiert)
Anz. Treffer           : 20.169
Anz. exportierte Zeilen: 10 (Exportoption)
Angezeigter Kontext    : 2 Sätze links, 2 Sätze rechts
Kontext umschließt     : 1. Wort des Treffers


A09      dass die Bauarbeiten spätestens 2009 hätten beginnen müssen.
A09  Euro, die wir an das Konsortium Sky Team hätten zahlen müssen.»
A09              an den Rand eines Atomkriegs hätten bringen können.
A09     ausser dass sie noch etwas mehr Wille hätte zeigen können.
A09 der Diepoldsauer Musik, Roland Stillhard, hat einspringen müssen:
A09  dem sich die Gemeinde in positivem Sinne habe präsentieren können.
A09 sie sich gegenüber der Gemeindeverwaltung habe rechtfertigen müssen.
A09  inwieweit (…) ihre Preise nach unten werden anpassen müssen.
A09 die leicht zu ähnlichen Gewaltakten hätte führen können, wie nach den
A09     – sich sonst kaum Weihnachtsgeschenke hätten leisten können.
```

Abb. 116: Auswahl aus den Konkordanzzeilen zur Abfrage Verbalkomplex: fin. Auxiliar – Inf. Vollverb – Inf. Modalverb (DeReKo)

Die Konkordanzen zeigen ein zufriedenstellendes Bild. Es handelt sich tatsächlich um Verbalkomplexe am Ende von Nebensätzen. Die Fehlerquote ist sehr gering.
Um die Verteilung dieser Verbalkomplexe in Texten einzelner Länder zu beobachten, muss jetzt die Länderansicht aktiviert werden.

Die Länderansicht erreicht man über den Befehl Ergebnisse → Länderansicht. Die Verteilung dieser Verbalkomplexe in deutschen (D), österreichischen (A) und schweizerischen (CH) Texten zeigt die Abb. 117. Die Ergebnisse sind nach der relativen Häufigkeit (berechnet pro eine Mio. Worte) gereiht.

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	331	1.385 pMW	331	1997	2009	**CH**
⊞	612	0.945 pMW	612	1998	2009	**D**
⊞	57	0.427 pMW	57	1999	2009	**A**
	1.000	0.980 pMW	1.000	1997	2009	3 Länder

Abb. 117: Länderansicht zur Abfrage Verbalkomplex: fin. Auxiliar–Inf. Vollverb–Inf. Modalverb (DeReKo)

Dieses Rechercheergebnis (Abb. 117) wird mit dem Ergebnis in der Abfrage 2 (Abb. 118 auf der folgenden Seite) verglichen.

Abfrage 2:
In dieser Abfrage wird nach der „österreichischen" Reihenfolge der Verben im Verbalkomplex gesucht, wie sie im Beispielsatz (2) auf Seite 155 erscheint. Der Anfang der Abfrage (die Archiv- und Korpuswahl sowie die Einstellung der Optionen) ist gleich wie in der Abfrage 1.

- **Suchanfrage** (Eingabe ins Suchfeld):
  Es wäre logisch die Suchanfrage an den Korpusmanager COSMAS II so zu formulieren, wie es in der Abfrage 1 der Fall war, nur mit getauschten Positionen. Also: Infinitiv des Vollverbs – Finites Hilfsverb - Infinitiv des Modalverbs am Satzende. „Übersetzt" in die COSMAS II-Sprache:
  MORPH(VRB inf v)·MORPH(VRB fin a)·MORPH(VRB inf m)·#IN(R) <s>

  Diese Anfrage ergibt jedoch keine Ergebnisse. Der Grund ist naheliegend: die Reihenfolge wird eben als „ungrammatisch" betrachtet und dies wirkt sich auch auf die Annotation des Korpus aus. Bei der Annotation wird nämlich jedem Wort sein Tag u.a. nach seiner Umgebung zugewiesen[44]. Deswegen muss man auf eine einfachere Eingabe ausweichen und statt Infinitiv modal (die letzte Position im Verbalkomplex) Infinitiv „beliebig" eingeben.

Die Eingabe lautet jetzt (**Achtung!** Leerzeichen an den richtigen Stellen hinzufügen!)

MORPH(VRB inf v)·MORPH(VRB fin a)·MORPH(VRB inf)·#IN(R)·<s> → Suchen

Die Abfrage bedeutet: Suche Infinitiv eines Vollverbs (VRB inf v), finites Hilfsverb (VRB fin a) und Infinitiv eines beliebigen Verbs (VRB inf) in dieser Reihenfolge und am Ende (#IN(R) = „rechtes Ende") eines Satzes (<s>).

Nachdem die Treffer geliefert worden sind, kann man die Länderansicht betrachten. Sie ist in der Abb. 118 auf der nächsten Seite zu sehen.

[44] Fein annotierte Korpora (z.B. deTenTen) können diese Abfrage beantworten und liefern zufriedenstellende Ergebnisse. Ihr Nachteil ist wiederum, dass sie oft keine Länderansicht ermöglichen.

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	184	1.377 pMW	184	1999	2009	**A**
⊞	81	0.339 pMW	81	1997	2009	**CH**
⊞	165	0.255 pMW	165	2002	2010	**D**
	430	0.421 pMW	430	1997	2010	3 Länder

Abb. 118: Länderansicht zur Abfrage Verbalkomplex: Inf. Vollverb – fin. Auxiliar – beliebiger Infinitiv

Die Ergebnisse der Verteilung dieser Verbalkomplexe in deutschen (D), österreichischen (A) und schweizerischen (CH) Texten sind in der Abb. 118 nach der relativen Häufigkeit (berechnet pro eine Mio. Worte) gereiht. Es geht deutlich hervor, dass die Mehrheit dieser Verbalkomplexe (sowohl absolut, als auch relativ gerechnet) in österreichischen Texten vorkommt.

Die KWICs bzw. Volltexte sollten nun näher nach einzelnen Ländern betrachtet werden. Ruft man die deutschen (Abb. 119) oder schweizerischen (Abb. 120) Konkordanzen ab, sieht man überwiegend solche Belege, die nicht in diese Problematik gehören, denn bei der Abfrage werden alle Abfolgen von drei Verben (mit hier definierten Eigenschaften) abgerufen, also auch diejenigen, die zwar hintereinander stehen, jedoch keinen Verbalkomplex bilden. Sie stehen zwar alle am Ende eines komplexen Satzes, jeweils eines oder zwei von diesen Verben gehören jedoch zum eingeschobenen Satz.

BRZ09	solch einer Tat zu	**vermindern, ist zuzuhören.**	Den Opfern,
HAZ09	dem künftigen Pensionär zu	**tun hatten, sagen,**	er hätte immer ein
HAZ09	lohnt. „Ich würde gern	**mithelfen, würde mitgehen**	und ihnen
HMP09	können gar nicht auf ein 0:0	**spielen, werden angreifen**	- mit
HMP09	auf den Turniersieg zu	**stürzen, war abzusehen.**	"Ich habe
RHZ09	wir Anfang nächsten Jahres	**veröffentlichen werden, sagen**	95,5

Abb. 119: Auswahl fehlerhafter Konkordanzen zur Abfrage Infinitiv Vollverb – finites Hilfsverb – beliebiger Infinitiv, Texte aus Deutschland

A08	nur noch, den Käfig, den ich jetzt wohl	**verschenken würde, auszumisten.**
A99	dämmerigen Kirche, sofern man sie dort	**antreffen würde, erkennen.**
A08	von Markus Sprenger (CVP), der in Pension	**gehen wird, antreten.**

Abb. 120: Auswahl fehlerhafter Konkordanzen zur Abfrage Infinitiv Vollverb – finites Hilfsverb – beliebiger Infinitiv, Texte aus der Schweiz

Das Problem der Überlappung der Sätze kann man damit umgehen, dass die Beistriche in der Abfrage negiert werden (Negationsoperator ist %). Dabei müssen auch die Klammern entsprechend eingesetzt werden.
Die Suchfeldeingabe sieht dann folgendermaßen aus (**Achtung!** Leerzeichen an den richtigen Stellen hinzufügen!):

((MORPH(VRB inf v) %+w1 ,) MORPH(VRB fin a) %+w1 ,) MORPH(VRB inf) #IN(R) <s>

Die Ergebnisse dieser Abfrage sind ziemlich eindeutig: Wenn in einem deutschen oder schweizerischen Text das Auxiliar an zweiter Stelle steht, dann nur in Verbalkomplexen mit dem Verb *lassen*, wie die Abb. 121 zeigt.

Deutsche und schweizerische Texte

(D)	sich in der Region etwas zu Schulden	**kommen haben lassen.**
(D)	der das „Apollo“ (…) zwölf Jahre lang	**leerstehen hat lassen**.
(D)	die Ablehnung einiger kleiner Parteien (…)	**aufhorchen hätte lassen.**
(Ch)	Emotionalität setzende Zeichensprache	**hineinfliessen hat lassen.**
(Ch)	und Flurin Caviezel die Fränzlis wieder	**aufleben haben lassen.**
(Ch)	von der Schweizerischen Gesellschaft für Hotelkredite	**prüfen hatte lassen.**

Abb. 121: Auswahl der Konkordanzen zur Abfrage Infinitiv Vollverb – finites Hilfsverb – beliebiger Infinitiv, ohne Beistriche (DeReKo, Tagged-T)

Die Belege aus österreichischen Texten (A) liefern ein anderes Bild: das Auxiliar an zweiter Stelle erscheint in den Verbalkomplexen nicht nur, wenn am Ende das Verb *lassen* steht, sondern auch wenn diese Position Modalverben einnehmen. Wahrnehmungsverben und andere Verben mit Infinitiv ohne *zu* (AcI-Verben) kommen aber äußerst selten vor. Es scheint, dass sich diese „österreichische“ Reihenfolge auf Verbalkomplexe mit einem Modalverb oder mit dem Verb *lassen* beschränkt. Zumindest in gesprochenen Texten sind andere AcI-Verben rar, dennoch belegt, wie der Abb. 122 zu entnehmen ist.

Österreichische Texte

NON09	hat man die Gewissheit, dass man	**weichen wird müssen.** "Wenn der Tunnel
X99	Schilling, die man über erhöhte Beiträge	**finanzieren habe müssen.**
BVZ09	Mannschaft, die jederzeit noch weiter	**zuschlagen hätte können.**
BVZ09	weil er sein Kind (…) in die zweisprachige Volksschule	**schicken hätte können.**
NON09	Bürger, die nur gemeinsam etwas	**bewegen hätten können.** Gewonnen hat
NON09	, da der Verein den Traktor gar nicht	**verkaufen hätte dürfen.**
BVZ08	, wer (…) in Traiskirchen um den heurigen Titel	**kämpfen wird dürfen.**
NON09	Kindergarten, dessen Bau schon längst	**beginnen hätte sollen.** "Wir haben einen
NON09	gemacht werden, die das Buntmetall	**verkaufen hätte sollen.** Dort erfuhren
BVZ08	und Ist. Links das iV-Center, wie es	**aussehen hätte sollen.** Rechts der
NON09	, dass der Schiri die beiden Streithanseln	**austauschen hätte lassen.** So erwartet
NON09	jene Person, die der Unternehmerin das Heroin	**zukommen hat lassen.**
X99	Spiele, die sich der Veranstalter für die jungen Gäste	**einfallen hatte lassen.**
NON08	dass sie einen (…) Mann mit "Kapuze und weißer Unterhose"	**weglaufen hat sehen**.

Abb.. 122: Auswahl der Konkordanzen zur Abfrage Infinitiv Vollverb – finites Hilfsverb – beliebiger Infinitiv, Texte aus Österreich

Vergleicht man die Abb. 119, 120 und 121 mit der Abb. 122, dann kommt klar zum Vorschein, dass die Ergebnisse der Länderansichten (Abb. 117, Seite 157 und 118, Seite 159) nicht trügen. Die Topographie der dreiteiligen Verbalkomplexe weist länderspezifische Abweichungen auf.

Ergänzend wird nun das Phänomen auch noch im InterCorp abgefragt.

Recherche im InterCorp

Die Abfrage lässt sich über den Abfragemodus CQL erstellen, allerdings ohne die Möglichkeit, die Ansicht nach Ursprungsland der Texte abzurufen[45].

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query type: CQL

Abfrage 1: „Deutsche" Reihenfolge: finites Hilfsverb - Infinitiv Vollverb – Infinitiv Modalverb

- CQL (Suchfenstereingabe):
  [tag="VAFIN.*"] [tag="VVINF.*"] [tag="VMINF.*"] [word="[;,\.\!]"] → Search

Diese Abfrage ist sowohl struktur- als auch zeichenmäßig eigentlich ident mit der **Abfrage 1 im DeReKo** (Seite 156). Sie bedeutet: Suche finites Hilfsverb (VAFIN), Infinitiv Vollverb (VVINF), Infinitiv Modalverben (VMINF) gefolgt von einem Satzendezeichen [;,\.\!].
Die Abfrage liefert fast 4.000 Belege.

Abfrage 2: „Österreichische" Reihenfolge: Infinitiv Vollverb – finites Hilfsverb - Infinitiv Modalverb. Im InterCorp kann (im Unterschied zum DeReKo) diese Abfrage ohne Einschränkung durchgeführt werden.

- CQL (Suchfenstereingabe):
  [tag="VVINF.*"] [tag="VAFIN.*"] [tag="VMINF.*"] [word="[;,\.\!]"] → Search

Das Ergebnis der Abfrage beträgt lediglich 27 Belege. Sie bestätigen jedoch die Ergebnisse der Recherche im DeReKo, wenn die Quellen näher angesehen werden: In der ersten Spalte ist der „technische" Titel (deswegen auf Tschechisch) der Datei angeführt, in dem der Beleg vorkommt.

	opus	links	KWIC
(1)	jelinek-pianistka	Er prophezeit, daß er sicher drei Tage nicht	**gehen wird können.**
(2)	Bachmannova-Povidky	diese Person, die auch noch Todesurteile	**unterzeichnen hatte müssen,**
(3)	simmel-a_jimmy_sel_za	Nur das Geschwätz, das er halt	**schreiben hat müssen,**
(4)	Frankova-DenikA_Franko	, weil er weiß , welche Opfer Mutter	**bringen hat müssen .**
(5)	_EUROPARL	dass man die (…) Katastrophe etwas stärker	**hervorheben hätte müssen.**
(6)	Grass-Sire_pole	» bei seinem Vortrag, den er ja doch nich	**halten wird können, (…) «**

Abb. 123: Konkordanzzeilen zur Abfrage Infinitiv Vollverb – finites Hilfsverb – Infinitiv Modalverb (InterCorp)

Die hier ausgewählten Beispiele sind aus Büchern österreichischer Autor/-innen (1-3), aus einer Übersetzung einer in München lebenden Übersetzerin[46] (4) und aus festgeschriebenen Reden im EU-Parlament (5), wo die Identität des Autors nicht feststellbar ist. Ein Beispiel (6) stammt aus einem Werk von Günter Grass.

[45] Die Option „Suche nach Ursprungsland der Texte" ist im InterCorp erst in Vorbereitung.
[46] Mirjam Pressler: Anne Frank Tagebuch http://www.mirjampressler.de/about/

Fazit:
Die Verbalkomplexe mit drei Verben haben in den österreichischen Texten eine andere Topographie (Reihenfolge) als in deutschen und schweizerischen Texten. Das Auxiliar erscheint in österreichischen Texten mehrheitlich in der mittleren Position. Dies gilt für Verbalkomplexe mit Modalverben, bedingt auch für andere AcI-Verben *hören* (7) und *sehen* (8)[47]:

(7) Ich fand es sehr heftig, ja schon, aber hatte es mir noch brutaler vorgestellt, so wie man die Leute **reden hat hören**.
(8) Frau Präsidentin , wir sind der Auffassung , dass die im Entwurf vorliegenden Schlussfolgerungen des Rates , die wir alle diese Woche **durchsickern haben sehen ,** eine sehr reale Gefahr für die Europäische Union bedeuten könnten .

Verbalkomplexe in deutschen und schweizerischen Texten haben gelegentlich auch die „österreichische" Reihenfolge, wenn an ihrem Ende das Verb *lassen* steht (9), (10):

(9) …, der sich selbst als einer der ersten **impfen hat lassen.**
(10) …, die bei mir die Alarmglocken **schrillen haben lassen.**

Wie sich die Verbalkomplexe mit anderen AcI-Verben verhalten, muss noch recherchiert werden.

Schlussbemerkung zur Studie 8
Mit dieser Studie entstehen (mindestens) zwei Fragen: 1) Ist die bestehende Annotation der Korpora ausreichend? 2) Warum findet ein so häufiges Phänomen, welches hier behandelt wurde, keine Resonanz in Grammatiken? Es ist nur zu hoffen, dass diese (und andere) Lücken in der Beschreibung der deutschen Sprache mit dem internationalen Projekt „Variantengrammatik des Standarddeutschen" (Dürscheid/ Elspaß/ Ziegler 2014) geschlossen werden.

[47] Belege aus dem deTenTen-Corpus (Web-Corpus an der Masaryk-Universität); Abfrage: CQL: [tag="V.*"] [lemma="haben"] [lemma="hören"] [word="[;,\.\!]"], bzw. [tag="V.*"] [lemma="haben"] [lemma="sehen"] [word="[;,\.\!]"]

Studie 9: Präposition *pro*

Auch diese Studie bleibt auf der syntaktischen Ebene und betrachtet dabei die deutsche Sprache aus plurizentrischer Sicht. Der Anlass dazu war ein Disput mit einem Kollegen über die Formulierung eines Prüfungstests. In der Punktevergabe hieß es: *pro richtiger Antwort 5 Punkte*. Diese Angabe hat er als grammatikalisch falsch angezeigt, mit dem Hinweis, es solle heißen *pro richtig**e** Antwort*, ich möge den Fehler korrigieren und das Testformular neu ausdrucken. Bauchgefühle täuschen, deswegen gibt es Nachschlagewerke – und Korpora.

Im Duden Universalwörterbuch (2006) steht folgender Eintrag:

> **pro** <Präp. mit Akk.> [lat. pro = vor, für, anstatt]:
> **1.** *jeweils, je, für (jede einzelne Person od. Sache):* p. Person [und Jahr]; 100 km p. Stunde; er rasiert sich einmal p. Tag.
> **2.** *[1]für* (1 b).

Im Österreichischen Wörterbuch (2001: 459) findet man keine explizite Angabe über den Kasus, den die Präposition *pro* regiert. Die Schwankung zwischen dem Akkusativ und Dativ wird jedoch am Verwendungsbeispiel angedeutet:

> **pro;** pro (per) Stück: für jedes Stück; pro anwesend**e(r)** Dame …

Im Variantenwörterbuch (2004) gibt es dazu keinen Eintrag.

Ob es sich um einen regionalen Unterschied oder um eine gesamtdeutsche Schwankung handelt, kann man an Daten solcher Korpora überprüfen, in denen die Markierung (Metadaten) über den Ursprung der Texte abrufbar ist. Dies ist der Fall im DWDS und im DeReKo. Alle modernen Texte (nach 1945) im DWDS stammen ausschließlich aus Deutschland, deswegen scheidet dieses Korpus für eine plurizentrische Recherche aus. Ob die Kasusschwankung auf regionaler Ebene betrachtet werden kann, überprüft man am besten im **DeReKo**. Gesucht wird nach der Präposition *pro*, die von einem Adjektiv in attributiver Funktion gefolgt wird.

- **Cosmas II$_{web}$** → Recherche →
- Archiv: TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger) →
- Korpus: TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T
- **Optionen**: **Suchmodalitäten:**
  - Groß- / Kleinschreibung beachten für 1. Zeichen — abwählen
  - Groß- / Kleinschreibung beachten für andere Zeichen — abwählen
  - **Expansionslisten anzeigen:** — abwählen
  - Übernehmen
  - **Ergebnispräsentation:**
    - Länderansicht — auswählen
  - Übernehmen
- **Suchanfrage** (Eingabe ins Suchfeld)**:**
  MORPH-Assistent → Adj. → attributiv — auswählen

Im Suchfeld erscheint: **MORPH(ADJ at)**, dann den Cursor vor MORPH positionieren und davor schreiben: pro·/+w1·.

Die Abfrage sieht folgendermaßen aus (Achtung! Leerzeichen an den richtigen Stellen hinzufügen!):

pro·/+w1·MORPH(ADJ·at) → Suchen

Die Abfrage bedeutet: Suche alle Formen *pro* und im Abstand von einem Wort rechts alle Adjektive in attributiver Position. Die Ergebnisse werden in Form einer Tabelle mit der Verteilung in Texten in einzelnen Ländern ausgegeben(Abb. 124). Man kann erkennen, dass die Kombination *pro* mit Adjektiv am häufigsten in schweizerischen Texten vorkommt.

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	491	3.675 pMW	457	1999	2009	A
⊞	996	4.168 pMW	926	1997	2009	CH
⊞	1.775	2.741 pMW	1.673	1998	2010	D

Abb. 124: Ergebnisse *pro*+Adj. attributiv, Länderansicht (DeReKo)

Zu erkennen ist dies an der Spalte „relative Häufigkeit" (**rel. Häuf.**), die das Vorkommen pro eine Mio. Worte angibt.

Eine direkte Ermittlung des Kasus ist im Archiv TAGGED-T nicht möglich[48], daher muss man die Kookkurrenzanalyse durchführen, erst dann aus den syntagmatischen Mustern die Kasusformen ableiten und von ihnen die Frequenzen in einzelnen Ländern abrufen.

- Kookkurrenzanalyse

  Einstellungen: Kontext:

  0 Wörter links 2 Wörter rechts

Die restliche Einstellung kann in der Standardeinstellung bleiben → Starten.
Die Berechnung der Kookkurrenzen (Belica 1995) liefert folgende syntagmatische Muster:

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
1	295	295	2626	**Kilometer**	81% pro gefahrenem|gefahrenen Kilometer
2	392	97	2031	**verkauftem**	83% pro verkauftem
3	486	94	2024	**mente**	85% pro mente "
4	493	7	1629	**gefahrenem**	100% pro gefahrenem
5	569	76	1324	**verkaufter**	73% pro verkaufter Karte
6	629	60	1263	**gelaufener**	81% pro gelaufener Runde
7	634	5	1191	**gefahrenen**	60% pro gefahrenen
8	761	127	1093	**Person**	85% pro erwachsene Person
9	762	1	1068	**gelaufenem**	100% pro gelaufenem
10	818	56	960	**angefangene**	87% pro angefangene Stunde
11	875	57	898	**laufendem**	84% pro laufendem Meter
12	930	55	875	**gelaufene**	83% pro gelaufene Runde

Abb. 125: Auswahl der Kookkurrenzen und syntagmatischer Muster zur Abfrage Präp. *pro* und Adjektiv (DeReKo, TAGGED-T)

[48] Die Abfrage nach Kasus über Tag ist nur im DeReKo-Korpus Tagged-M möglich. Dieses Korpus besteht allerdings nur aus bundesdeutschen Texten (vor allem Mannheimer Morgen und Der Spiegel – siehe Textorganisation unter COSMAS II → Korpora).

Gleich das erste Syntagma (*pro gefahrenem/ gefahrenen Kilometer*) deutet auf die Kasusschwankung hin. Ob es signifikante Unterschiede im regionalen Bereich gibt, lässt sich mit der nächsten Abfrage feststellen.

Es ist ratsam diese im Archiv W - Archiv der geschriebenen Sprache durchzuführen, da dieses wesentlich größer ist:

- Archiv: W - Archiv der geschriebenen Sprache →
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: (keine Änderung notwendig)
- **Suchanfrage** (Eingabe ins Suchfeld)**:** **pro·gefahrenem·Kilometer** → Suchen
  dann **pro·gefahrenem·Kilometer** → Suchen

Auch die weitere(n) Suchanfrage(n) nach den syntagmatischen Mustern aus der Abb. 125 sollte(n) mit derselben Einstellung erfolgen, um die Validität der Daten zu gewährleisten.

Die Ergebnisse der Abfragen in den Abb. 126 und 127 deuten darauf hin, dass die Kasusschwankung der Präposition *pro* tatsächlich regional bedingt sein kann:

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	64	**0.0939 pMW**	57	1992	2013	A
⊞	158	**0.0504 pMW**	152	1987	2013	D
⊞	10	**0.0216 pMW**	10	1997	2013	CH
	232	0.0542 pMW	219	1987	2013	3 Länder

Abb. 126: Ergebnis der Abfrage *pro gefahrenem Kilometer*, Länderansicht, sortiert nach rel. Häufigkeit absteigend (DeReKo)

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	45	0.0971 pM	43	1996	2013	CH
⊞	125	0.0399 pM	119	1994	2012	D
⊞	25	0.0367 pM	24	1992	2013	A
	195	0.0456 pM	186	1992	2013	3 Länder

Abb. 127: Ergebnis der Abfrage *pro gefahrenen Kilometer*, Länderansicht, sortiert nach rel. Häufigkeit absteigend (DeReKo)

Als Kontrollbeispiel wurde noch das syntagmatische Muster *pro gelaufener Runde*/ *pro gelaufene Runde* abgefragt. Auch diese Ergebnisse zeigen dieselben Verhältnisse der Aufteilung von Treffern:

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	29	0.0426 pMW	28	2000	2013	A
⊞	70	0.0223 pMW	70	2001	2012	D
⊞	2	0.0043 pMW	2	2007	2011	CH
	101	0.0236 pMW	100	2000	2013	3 Länder

Abb. 128: Ergebnis der Abfrage *pro gelaufener Runde*, Länderansicht, sortiert nach rel. Häufigkeit absteigend (DeReKo)

	Treffer	rel. Häuf.	Texte	von	bis	Land
⊞	21	0.0453 pM	18	1998	2013	CH
⊞	48	0.0153 pM	48	2004	2013	D
⊞	4	0.0059 pM	4	2008	2013	A
	73	0.0171 pM	70	1998	2013	3 Länder

Abb. 129: Ergebnis der Abfrage *pro gelaufene Runde*, Länderansicht, sortiert nach rel. Häufigkeit absteigend (DeReKo)

Zu diesen Ergebnissen gelangt man relativ rasch. (Man bedenke, wie viele Arbeitstage ein Mensch nur für eine dieser Recherchen brauchen würde.) In weiterer Folge wurden noch diese Abfragen durchgeführt: *pro angefangener*/ *pro angefangene*, *pro verkaufter*/ *pro verkaufte, pro verkauftem, pro verkauftes*. Sie alle bestätigten die Ergebnisse in den Abb. 126–129.

Fazit:
Die Ergebnisse zeigen ein plastisches Bild über den Gebrauch von *pro* + Dativ/ Akkusativ: In schweizerischen Texten wird fast ausschließlich Akkusativ verwendet, in österreichischen Texten wiederum in den meisten Fällen Dativ, die Akkusativform ist eine Seltenheit. Und in den deutschen Texten scheint dieses Phänomen als eine regelrechte Doppelform mit leichter Tendenz zum Dativ zu sein. Ob dies auch (innerhalb Deutschlands) regionalbeding ist oder ob textsortenspezifische Faktoren auch eine Rolle spielen, muss für weitere Untersuchungen offen bleiben. Auf jeden Fall ist mit dieser Studie ein weiteres Steinchen in das plurizentrische Mosaikbild der deutschen Sprache gelegt worden.

Kollokationen, Phraseme und Lexik

Ein reflektierter Umgang mit einer (Fremd-)Sprache soll immer die Frage aufkommen lassen: was ist in der Sprache üblich, was ist eher selten; was ist wichtig zu beherrschen, was ist weniger wichtig; welche Mittel braucht man um das Kommunikationsziel zu erreichen. Diese Mittel können im Kopf grundsätzlich über zwei Wege abgerufen werden: 1) als Ergebnis eines grammatisch-lexikalischen Prozesses, oder 2) als „Fertigteile" (Chunks), die als modifizierbare Einheiten gespeichert sind. Im Fremdsprachenunterricht scheint es sinnvoll, beide Wege zu kombinieren, wie Swan (2006) betonte. Die Überlegungen, welche Aspekte der Grammatik hervorgehoben werden sollen, welche Lexik und welche Fertigteile (Chunks) primär vermittelt werden sollen, hängen eng mit dem Ziel des Unterrichts zusammen. Wegen ihrer Breite kann auf sie nicht im Einzelnen eingegangen werden.

In diesem Kapitel werden einige wenige Fallbeispiele präsentiert, in denen sich mehrere sprachliche Ebenen und Aspekte verbinden: paradigmatische und syntagmatische Beziehungen, feste und freie Verbindungen, diachrone und synchrone Betrachtung der Lexik. Die ersten zwei Dichotomien sind stark mit der Phraseologie verbunden. Ohne an dem „Terminologiekrieg um die Besetzung des linguistischen Terminus *Kollokation*" (Ďurčo et al. 2010: 7) teilzunehmen[49], werden hier Beispiele gezeigt, wie Wortverbindungen mithilfe Korpora aufgedeckt werden können.

Studie 10: Kollokationen, syntagmatische Muster, Chunks

Die Frage danach, was in der Sprache typisch ist, was eher nicht, betrifft im nicht geringeren Maße die Wortverbindungen. Diese können mithilfe der Korpusmanager errechnet und elegant mit einer anderen Sprache verglichen werden, wie es z.B. für feste Wortverbindungen im deutsch-slowakischen Kontrast Ďurčo et al. (2010) gemacht haben.
Die Frage ist nun: Wie kann man die Wortverbindungen aus dem Korpus abrufen und so aufdecken?

Fragestellung 1: *Was sind die am meisten frequentierten Wörter in der Sprache und wie ermittelt man die häufigsten Verbindungen, in denen sie vorkommen?*

Fürs Deutsche wurde eine Liste der 100.000 häufigsten Wortformen und 40.000 Grundformen (DeReWo 2009) erstellt. Diese Liste wird folgend interpretiert: diejenigen Wörter, die oben stehen haben die höchste Frequenz, sie waren am häufigsten in Texten des DeReKo 2009 vertreten. Es ist daher sehr wahrscheinlich, dass sie allgemein in den meisten deutschen Texten vorkommen. Am Anfang dieser Liste stehen synsemantische Relationswörter, die in fast jedem Text vorkommen (Artikel, Präpositionen und Konjunktionen, Hilfsverben). Vergleicht man die jeweils zwanzig häufigsten Wörter in der Liste DeReWo (2009) mit der Word List aus InterCorp_de, Kernkorpus (Dovalil/ Káňa/ Peloušková et al. 2013), stellt man fest, dass sich diese grob entsprechen:

[49] Dazu im Detail: Hausmann 2003: 320-321.

Die zwanzig häufigsten Wortformen nach DeReWo (2009), alphabetisch:

auf, auch, das, dem, den, der, des, die, ein, eine, für, im, in, ist, mit, nicht, sich, und, von, zu

Die zwanzig häufigsten Wortformen im InterCorp_de (core) (2013), alphabetisch:

auf, das, dem, den, der, die, ein, er, es, ich, in, ist, mit, nicht, sie, sich, und, von, war, zu

Es sind (wenig überraschend) alles Synsemantika („Systemwörter"). Erst auf weiteren Frequenzpositionen stehen autosemantische Wörter[50]. Diese „entsprechen" dem Inhalt des Korpus. (Autosemantika werden nämlich nach Thema des Textes häufig/ weniger häufig/ gar nicht verwendet.) Auch hier gibt es Übereinstimmungen. Vergleicht man die ersten 200 häufigsten Lemmata im DeReWo und Intercorp_de (core), stellt man fest, dass die folgenden Wörter in beiden Korpora (obwohl in jedem Korpus andere Texte gespeichert sind und die Korpora unterschiedlich aufgebaut werden) am häufigsten vorkommen:

beginnen, bleiben, bringen, Ende, fahren, fragen, Frau, geben, gehen, groß, gut, halten, Haus, heißen, Jahr, jung, Kind, klein, kommen, kurz, lang, leben, liegen, legen, Mann, Mensch, sagen, sehen, Seite, schön, stehen, stellen, Tag, Weg, weit, Welt, zeigen, Zeit.

Diese Wörter bilden den „härtesten" Kern des deutschen Wortschatzes[51].
Als Beispiel wird in einem nächsten Schritt das Wort *Jahr* in seinen typischen Chunks ermittelt. Es ist übrigens das allerhäufigste autosemantische Wort gleich in mehreren Korpora: **DeReKo**, deTenTen, deWaC. Allgemein ist es im geschriebenen Deutsch also sehr gebräuchlich. Um auch die gesprochene Sprache annähernd einzubeziehen, wurde die Abfrage im DeReKo auf belletristische Texte eingeschränkt (Belletristik liegt am nähersten der gesprochenen Sprache wegen deren Nachahmung in der direkten, teilweise auch indirekten Rede). Dazu wurde ein Sub-Korpus aus zugänglichen belletristischen Texten (über 16 Mio. laufender Wörter) erstellt.

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache → Subkorpus erstellen
- Korpus → Korpusverwaltung → Benutzerdefinierte Korpora → Definieren
  gewünschte Texte in die rechte Spalte verschieben
  Name (Korpus benennen) → übernehmen → Speichern
- **Optionen**: Standardeinstellung
- **Suchanfrage** (Eingabe ins Suchfeld)**: &Jahr** → Suchen

Nachdem die Ergebnisse erschienen sind, wird die Kookkurrenzanalyse durchgeführt:

- Kookkurrenzanalyse
  Einstellungen: → Zurücksetzen (Standardeinstellung)
  Funktionswörter ignorieren abwählen → Starten

In Kürze erscheint ein Bild (Abb. 110) mit Kookkurrenzpartnern/ Kollokatoren (hier Kookkurrenzen genannt) und entsprechenden syntagmatischen Mustern, in denen sie vorkommen. Es zahlt sich aus, die syntagmatischen Muster durchzugehen. Jede/-r, der/die in

[50] Zur Definition von Auto- und Synsemantika siehe Fachlexikon DaF/ DaZ (2010: 22, 328).
[51] Die Ergebnisse wurden auch an Daten anderer großer Korpora überprüft (deTenTen und deWac).

der deutschen Sprache täglich lebt, stellt fest, dass viele von diesen Mustern tatsächlich „unglaublich vertraut“ klingen.

#	Total	Anzahl	Autofokus von	Autofokus bis	LLR	Kookkurrenzen	syntagmatische Muster	
1	3	3	1	1	94	**alt** dreißig	100%	dreißig Jahre alt und ...
2	71	3	-1	-1	71	**dreißig** lang	100%	dreißig Jahre lang
3	110	22	-1	-1	69	**zwanzig**	59%	vor zwanzig Jahren
4	117	7	-1	-1	41	**siebzehn**	57%	siebzehn Jahre
5	123	6	-1	-1	38	**fünfundzwanzig**	50%	fünfundzwanzig Jahre
6	139	16	-1	-1	34	**halbes**	93%	ein halbes Jahr
7	142	3	-1	-1	30	**sechsundzwanzig**	100%	sechsundzwanzig Jahre
8	152	10	-1	-1	26	**vierzig**	50%	vierzig [...] Jahre
9	157	5	-1	-1	25	**achtzehn**	80%	achtzehn Jahren
10	160	3	-1	-1	24	**einundzwanzig**	100%	einundzwanzig Jahre
11	165	5	-1	-1	19	**sechzig**	100%	sechzig Jahre
12	177	12	1	1	18	**älter**	75%	Jahre älter als ...
13	185	8	1	1	18	**jünger**	50%	Jahre jünger
14	187	2	1	3	17	**PGN**	100%	Jahr [zu Jahr] PGN
15	192	5	-1	-1	15	**dreizehn**	100%	dreizehn Jahre alt ...
16	196	4	-1	-1	15	**zwanziger**	50%	zwanziger Jahre
17	198	2	-5	-5	14	**Kontaktanzeige**	100%	Kontaktanzeige im folgenden J.
18	211	13	-1	-1	12	**halben**	100%	seit|vor einem halben Jahr
19	215	4	-1	-1	10	**vierzehn**	50%	vierzehn Jahre
20	217	2	2	4	10	**Seydlitz**	50%	Jahre ... Seydlitz
21	219	2	-1	-1	10	**vierhundert**	50%	vierhundert Jahre
22	221	2	-5	4	10	**Mäsjuh**	50%	Jahres ... Mäsjuh
23	223	2	-1	-1	9	**Fünfzehn**	100%	Fünfzehn Jahre
24	228	5	-1	-1	9	**fünfzehn**	80%	fünfzehn [...] Jahre
25	231	3	2	4	8	**Sils**	66%	Jahr nach Sils
26	235	4	-1	-1	8	**sechziger**	50%	sechziger Jahre
27	240	5	-1	-1	8	**fünfzig**	80%	fünfzig [...] Jahre
28	242	2	3	4	7	**zugebracht**	100%	Jahre im ... zugebracht
29	253	11	-2	-2	6	**Laufe**	90%	im Laufe der Jahre
30	256	3	-2	-1	6	**sechzehn**	100%	sechzehn [...] Jahre alt ...
31	259	3	-1	-1	6	**fünfziger**	100%	fünfziger Jahren
32	263	4	-4	-4	6	**zarten**	75%	im zarten Alter von ... Jahren
33	266	3	-1	-1	5	**Letztes**	100%	Letztes Jahr
34	269	3	-3	-1	5	**siebzig**	100%	siebzig [oder ...] Jahre
35	271	2	-2	-2	5	**hab's**	100%	hab's ... Jahre
36	273	2	2	2	4	**aufgebraucht**	50%	Jahre ... aufgebraucht
37	306	33	1	1	4	**lang**	60%	Jahre lang
38	312	6	1	1	3	**vergangen**	66%	Jahre [waren] vergangen
39	314	2	-2	-1	3	**tausenden**	100%	tausenden [...] Jahren
40	316	2	-1	3	3	**unbeschwerten**	50%	Jahre ... unbeschwerten
41	325	9	-1	-1	2	**hundert**	55%	hundert Jahren
42	327	2	4	4	2	**solltest**	100%	Jahre vor dir und solltest
43	346	19	-1	-1	2	**jedes**	84%	jedes Jahr
44	348	2	-1	-1	2	**1/2**	50%	1/2 Jahren
45	350	2	4	4	2	**wunderschöne**	100%	Jahren ... wunderschöne
46	352	2	2	2	2	**aufbewahrt**	50%	Jahre ... aufbewahrt
47	356	4	-2	-2	2	**deinen**	75%	deinen ... Jahren
48	361	5	-1	-1	2	**ganzes**	80%	ein ganzes Jahr
49	363	2	4	5	2	**Feinden**	100%	Jahre ... Welt ... Feinden
50	365	2	-3	-3	1	**litt**	100%	litt ... Jahren
51	367	2	2	3	1	**Voraus**	100%	Jahr im Voraus
52	373	6	4	4	1	**gelebt**	83%	Jahre [...] gelebt
53	375	2	-2	2	1	**einhundert**	50%	Jahr ... einhundert
54	377	2	2	5	1	**siebenmal**	100%	Jahre ... oder ... siebenmal
55	379	2	1	1	1	**gedauert**	100%	Jahre gedauert
56	383	4	-1	-1	1	**nächstes**	100%	nächstes Jahr
57	385	2	-1	-1	1	**achtziger**	100%	achtziger Jahre
58	388	3	-1	-1	1	**Nächstes**	100%	Nächstes [...] Jahr
59	391	3	-3	-2	1	**mochte**	66%	mochte ... Jahre
60	393	2	-1	-1	1	**siebziger**	50%	siebziger Jahre

Abb. 130: Kookkurrenzen und syntagmatische Muster zu *Jahr* (DeReKo)

Aus den syntagmatischen Mustern kann man einige Chunks ableiten (Tab. 26), die gefühlsmäßig auch in der gesprochenen Sprache besonders häufig vorkommen. Sie sollten auch vorrangig vermittelt werden.

X Jahre alt/ lang	*seit | vor einem halben Jahr*
vor X Jahren	*im vergangenen Jahr*
ein halbes Jahr	*jedes Jahr*
Jahre älter als	*nächstes Jahr*

Tab. 26: Chunks aus syntagmatischen Mustern mit Lemma *Jahr*

In den syntagmatischen Mustern sind natürlich auch phraseologische Einheiten zu finden:
im Laufe der Jahre
im zarten Alter von ... Jahren

Die Konkordanzen zu jedem berechneten Muster sind durch den Klick auf die entsprechende Zeile abrufbar.

Würde man im ganzen öffentlichen Korpus recherchieren, wären auch diese syntagmatischen Muster relevant:
freiwilliges (soziales/ ökologisches) Jahr
zu … Jahren Haft verurteilt

Sie sind jedoch nur für Sachtexte typisch. Ihre Frequenz in der gesprochenen Sprache kann mithilfe bestehender Korpora nicht überprüft werden.

Fazit:
Die häufigsten Wörter im Deutschen sind Synsemantika: Artikelwörter *der, die, das, den, dem, des, ein(e)*, gefolgt von Präpositionen (*in, für* etc.), dem grammatikalischen Partikel *zu* u.a.

Aus den autosemantischen Wörtern sind es:
Substantive: *Ende, Frau, Haus, Jahr, Kind, Mann, Mensch, Seite, Tag, Weg, Welt, Zeit*;
Adjektive: *groß, gut, jung, klein, kurz, lang, schön, weit*;
Verben: *beginnen, bleiben, bringen, fahren, fragen, geben, gehen, halten, heißen, kommen, leben, legen, liegen, sagen, sehen, stehen, stellen, zeigen*. (Es ist jedoch anzunehmen, dass die meisten dieser Verben desemantisiert in verbo-nominalen Phrasen vorkommen, deswegen haben sie sehr hohe Frequenz.)

Auf der Frequenzliste der häufigsten deutschen Wörter stehen zwischen den synsemantischen und autosemantischen Wörtern eine Reihe deiktischer Wörter (*er, heute, hier, ich, Sie/sie, mein, da*) und alle Hilfs- und Modalverben.

Autosemantische Wörter mit der höchsten Frequenz sollten als erste vermittelt werden. Die häufigsten Phrasen und Chunks, in denen sie vorkommen, deckt die Kookkurrenz/Kollokationsanalyse auf.

Man kann die Abfrage um ein Element erweitern (Fragestellung 2) oder gleich nach einem komplexeren Element wie Phrase, Chunk, mehrgliedriges Satzglied (Fragestellung 3) suchen. Beide Abfragen werden ähnlich durchgeführt.

Fragestellung 2: *In welchen semantischen Korrelationen stehen die Wörter Kind und Schule?* Für Lernende kann die Frage auch einfacher formuliert werden: *Was können Kinder mit Schule „tun"?*

Einige Phrasen oder Chunks können direkt aus den Konkordanzen festgestellt werden. Sogar wenig fortgeschrittene Lerner/-innen können aus vorsichtig ausgewählten Konkordanzzeilen eine induktive Aufgabe selbst lösen.

Die Recherche kann im beliebigen Korpus durchgeführt werden. Im Folgenden werden wieder belletristische Texte des **DeReKo** verwendet.

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache → Subkorpus Belletristik (Erstellung siehe Fragestellung 1)
- **Optionen**: Standardeinstellung
- **Suchanfrage** (Eingabe ins Suchfeld)**: &Schule /s0 &Kind** → Suchen

Die Abfrage bedeutet: Suche Lemmata *Schule* und *Kind* in einem Satz, die Reihenfolge im Satz ist beliebig. Als Ergebnis bekommt man ein paar Dutzend Konkordanzen, wie die Abb. 131 zeigt.

DIV	Frühe wieder zur Arbeit, während die **Kinder** der reichen Eltern zur **Schule** gehen
DIV	Kilometer entfernte Nachbardorf, dessen **Kinder** die neue **Schule** nicht besuchen sollen,
DIV	das Leben in der Dorfgemeinschaft: Die **Kinder** gehen von morgens 6 Uhr bis zum
DIV	warteten vor der Tür. Man wollte die **Kinder** schließlich daran gewöhnen, dass sie
DIV	ihre Tochter davon überzeugte, dass **Kinder** in die **Schule** gehen müssen. Wie sie das
DIV	Karla davon überzeugt, dass jedes **Kind,** was sie natürlich mit einschloss, zur
DIV	das war dann bitter. Früher mussten die **Kinder** von Möllersgrund und Springen bis nach
DIV	ich nicht. Zu der Zeit spielten wir **Kinder** nach der **Schule** oft am Waldrand, dort
DIV	werden, es war halt Krieg. Unsere **Schule** bekam dann auch öfters fremde **Kinder,**
DIV	zu trinken, und eines Tages wurden die **Kinder** von der **Schule** weg abgeholt und ins
DIV	ruhigen Augen stillstehend wie ein **Kind** in der **Schule,** aller Hohn und Spott wird
WAM	war. Also wahrscheinlich ein Dentist. Die **Kinder** promovierter Ärzte hatten sich in der

Abb. 131: Auswahl der Konkordanzen zur Abfrage Lemma *Schule* und Lemma *Kind* in einem Satz (DeReKo)

Aus diesen (nur einigen wenigen ausgewählten) Konkordanzen können die Lerner folgende Schlussfolgerungen ziehen: was im Satz wichtig ist, welche Konstruktionen gebräuchlich sind, welche „Korrelationen" es zwischen *Schule – Kind* geben kann.

Kind(er) in der Subjektrolle	*Kind(er)* in Rolle eines Objektes
Kinder gehen zur/ in die Schule *Kinder müssen in die Schule gehen* *Kinder besuchen die Schule* *Kinder machen sich lustig über..*	*Unsere Schule bekam dann auch öfters fremde Kinder* *Kinder werden abgeholt*

Tab. 27: Das Wort *Kind* und seine Rollen im Satz mit *Schule*

Am häufigsten steht *Kind* (meistens im Plural) in der Rolle eines Subjekts/ Agens (Tab. 27 links) oder Objekts des Geschehens (Tab. 27 rechts).

Fragestellung 3: *Wie sucht man nach mehrteiligen Prädikaten und entdeckt die Rollen um das Prädikat herum?*

Im syntaktischen Zentrum des Satzes stehen einfache oder komplexe[52] Verben, die den Kern des Prädikats bilden. Prädikate sind auch entweder (1) einfach (nur Vollverben in finiter Form) oder (2) komplex (finite Form eines „entleerten" Hilfsverbs[53] in Verbindung mit einem „lexikalischen Kern" (Duden - Grammatik 2005: 422) des Prädikats).

(1) Die Abfrage nach einfachen Prädikaten erfolgt in jedem Korpus über die Funktion Lemma (siehe Fallbeispiel *Jahr* auf Seite 167). Besteht das einfache Prädikat aus einem Verb mit Zusatz, dann entspricht die Abfrage dem Vorgang auf der Seite 151 und 152.
(2) Komplexe Prädikate (hier das Beispiel *bekannt sein*) können in **DeReKo** und in **InterCorp** wie folgt abgefragt werden.

Recherche im DeReKo

- **Cosmas II** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: Standardeinstellung
- **Suchanfrage** (Eingabe ins Suchfeld)**: &sein /s0 bekannt** → Suchen

Die Abfrage bedeutet: Suche Lemma *sein* und das Wort *bekannt* in einem Satz.

Recherche im InterCorp

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query type: CQL
- CQL (Suchfeldeingabe)**:**

  <s/> containing [lemma="sein"] containing [word="bekannt"] → Search

Die Abfrage bedeutet: Suche Sätze, die das Lemma *sein* und das Wort *bekannt* beinhalten.

Die Konkordanzen können recht unübersichtlich sein, denn der Korpusmanager gibt alle Sätze wieder, in denen diese zwei Wörter vorkommen: man sieht also (auf dem Bildschirm) rot! Es besteht aber eine alternative Abfrage:

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query type: Word Form (Match case auswählen empfohlen)
- Word Form (Suchfeldeingabe)**: bekannt**

Nachdem die Konkordanzzeilen erschienen sind, führt man die Kollokationsanalyse durch:

- Collocation → Custom…

---

[52] Einfache Verben werden durch lexikalische Simplizia (d.h. lexikalisch nicht weiter zerlegbare Worte) realisiert (*gehen, laufen, schreiben …*).
Komplexe Verben sind entweder verbale Ableitungen (*beschreiben, verlaufen, zerkochen …*) oder verbale Komposita/ Verben mit Zusätzen (*abschreiben, weglaufen, kennenlernen …*).

[53] Hilfsverb oder Auxiliar wird hier als Sammelbegriff für Hilfsverben im traditionellen Sinne (*haben, sein, werden*) verwendet, weiterhin für Modalverben, Prädikativ(Kopula)- und Funktionsverben. (Hierzu: Duden Grammatik 2005: 420-423.)

Einstellung der Kollokationsanalyse (Collocations candidates):

- Attribute: Lemma
- In the range from: -5 to 5
- Minimum frequency in corpus: 10 (spielt keine Rolle für diese Abfrage)
- Minimum frequency in given range: 10 (spielt keine Rolle für diese Abfrage)

Make Candidate List

Unter den signifikanten Kollokationspartnern findet man die Verben *geben*, *werden*, *machen* und auch das gesuchte *sein*. Durch einen Klick auf p (positiver Filter) erscheinen die Konkordanzen mit den automatisch errechneten Kollokationspartnern (Abb. 132). Unter ihnen sind auch fehlerhaft errechnete Kollokationen wie diese:

> »Das (…) Fischereiministerium gibt heute **bekannt**, dass -« »**Ist** doch nicht zu fassen! … «
> Valerie hatte die Männer miteinander **bekannt** gemacht, sie **waren** nun sehr häufig zusammengetroffen -

Die Fehlerquote liegt jedoch unter 5%.

Ich denke, es **ist** uns allen	**bekannt**	, daß wir (…) vor einem Problem stehen.
(…) erwartet werden kann, daß sie ihm	**bekannt**	**sind**, und …
, daß dem Leistenden die Eröffnung nicht	**bekannt**	**war** .
dann hatte er erwähnt, dass Crandle dafür	**bekannt**	**war**, nicht viele (…) zu geben…
Oder wenn ihr	**bekannt**	**gewesen wäre**, dass ich keine, …

Abb. 132: Auswahl der Konkordanzen zur Abfrage *bekannt sein* in einem Satz (InterCorp)

Nach diesem Vorgang lassen sich alle mehrteiligen Prädikate abrufen. Aus den Kollokationsanalysen könne die typischen Prädikationen (= Basisgerüst jeder minimalen sinnvollen Aussage) aufgedeckt werden, wie auch der folgende Schritt zeigt.
Als Prädikat steht ein Verb mit einen Zusatz: *anziehen*.

Sucht man nach dem typischen Umfeld von einem Verb mit Zusatz, ist der Vorgang ähnlich der vorherigen Fragestellung 2. Man soll jedoch das ganze Paradigma des Verbs *anziehen* abfragen. Dementsprechend müssen auch alle Formen mit dem abgetrennten Zusatz abgerufen werden (vgl. auch Studie 7)

Recherche im DeReKo

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: Standardeinstellung
- **Suchanfrage** (Eingabe ins Suchfeld)**:**

&anziehen·oder·(&ziehen·/s0·an·#IN(R) <s>) → Suchen

Die Abfrage bedeutet: Suche alle Formen des Verbs *anziehen* (&anziehen) oder (oder) alle Formen des Verbs *ziehen*, wobei am Ende (#IN(R)) desselben Satzes (/s0) das Element *an* (an) vorkommt.

An den Ergebnissen (KWICs) wird wieder die Kookkurrenzanalyse durchgeführt:

- Kookkurrenzanalyse
  Einstellungen: → Zurücksetzen (Standardeinstellung) → Starten

Das Ergebnis der Berechnung wird wieder in Form einer Tabelle (Abb. 133) mit syntagmatischen Mustern präsentiert.

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
1	1077	1077	9649	**warm**	80% sich warm [...] anziehen
2	1608	531	6867	**magisch**	29% magisch [...] angezogen
3	3118	1510	6776	**Besucher**	19% zog [viele] Besucher an
4	3382	264	2644	**Zügel**	29% die Zügel [...] angezogen
5	4187	805	2509	**viele**	20% zog [...] viele ... an
6	4456	269	2182	**Massen**	27% zieht [die] Massen an
7	4457	1	2141	**Handbremse** vergessen	100% vergessen Handbremse anzuziehen
8	4911	256	1693	**Konjunktur**	32% wenn die Konjunktur [wieder] anzieht
9	5397	486	1436	**Publikum**	26% Publikum [...] anziehen
10	5628	231	1421	**Touristen**	27% und zieht Touristen an
11	5645	17	1328	**Schuhe** feste	70% feste Schuhe [...] anziehen
14	6388	103	947	**Schraube**	41% die Schraube [...] angezogen
15	6463	75	914	**Daumenschrauben**	33% die Daumenschrauben [...] angezogen
16	6580	117	903	**Handschuhe**	35% Handschuhe [...] anziehen
19	6938	3	841	**Kleidung** bequeme warme	100% warme bequeme Kleidung a.
34	8860	66	487	**Schrauben**	42% die Schrauben [...] angezogen
37	9058	77	453	**Jacke**	37% eine|die Jacke [...] anziehen
43	10212	55	358	**Pullover**	32% einen ... Pullover anziehen

Abb. 133: Auswahl der Kookkurrenzen zu *anziehen* (DeReKo)

Aus den syntagmatischen Mustern geht deutlich die Diversität der Lesarten dieses Verbs hervor: *sich warm anziehen*, *Schuhe/ Handschuhe/Kleider/ (warme, bequeme) Kleidung a~, Handbremse a~*; in übertragenem Sinne: *Besucher*, *Massen*, *Preise*, *Schrauben*, (hoffentlich auch nur metaphorisch) *Daumenschrauben a~*; *jmd/etw. zieht (jmdn./etw.) an* (*ein Künstler zieht viele Besucher/ großes Publikum an*, *die Konjunktur zieht (wieder) an*).

Dieselbe Abfrage kann auch im InterCorp durchgeführt werden.

Recherche im InterCorp

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Query type: CQL
- CQL (Suchfeldeingabe)**:**
  [lemma="anziehen"]|[lemma="ziehen"] []* [word="an"] within <s id=".*"/> → Search

Die Abfrage bedeutet: Suche alle Formen von *anziehen* oder (|) *ziehen*, davon rechts im beliebigen Abstand das Element *an*, allerdings nur bis zur Grenze des Satzes.
Auch im InterCorp wird jetzt die Kollokationsanalyse durchgeführt:

- Collocation → Custom…

Einstellung der Kollokationsanalyse (Collocations candidates):

- Attribute: Lemma
- In the range from: -5 to 5
- Minimum frequency in corpus: 5 (empfohlen)
- Minimum frequency in given range: 3 (empfohlen)
  Make Candidate List

Das Ergebnis bekommt man nicht in Form der syntagmatischen Muster (im Unterschied zum COSMAS II), sondern als einzelne Lemmata oder Wortformen (je nach Einstellung). Signifikante Kollokationspartner bzw. Nachbarn des Wortes sind in der Abb. 134 angeführt.

			Freq	logDice
1.	p/n	Kleid	49	8.211
2.	p/n	vorbei	43	7.615
3.	p/n	heran	28	7.595
4.	p/n	Hemd	25	7.540
5.	p/n	Schuh	28	7.529
6.	p/n	Brust	32	7.458
7.	p/n	Uniform	22	7.403
8.	p/n	waschen	22	7.374
9.	p/n	Bein	37	7.336
10.	p/n	Hose	21	7.274

			Freq	logDice
11.	p/n	Mantel	20	7.189
12.	p/n	Stiefel	16	7.094
13.	p/n	Betracht	23	7.044
14.	p/n	Haar	32	7.043
15.	p/n	Schluß	35	7.034
16.	p/n	duschen	13	6.980
17.	p/n	ausländisch	22	6.979
18.	p/n	Bett	32	6.938
19.	p/n	Rock	15	6.876
20.	p/n	Knie	17	6.867

Abb. 134: Auswahl aus signifikanten Kollokationspartnern zum Verb *anziehen* (InterCorp)

Über den positiven Filter (klicken auf p) gelangt man zu den Konkordanzen, in denen das Verb *anziehen* und der entsprechende Kollokationspartner in einem Satz stehen. Diese decken interessante Prädikationen auf. Einige von ihnen wurden hier ausgewählt. Die eher „unerwarteten" Kollokationspartner sind unterstrichen (2-9).

(1) Zwei Stunden lang **zog ich an der TV-Kamera einen Drahtzaun vorbei** , um die Illusion der Fahrt vollkommen zu machen
(2) … und **zog die Faust** mit ganzer Kraft **an die Brust**, …
(3) Instinktiv **zog sie die Decke an die Brust**.
(4) aber sie trank rasch einen Schluck, während **sie sich** eilig **wusch** und **anzog.**
(5) Morgens **war er** lange vor mir wach, **gewaschen** und **angezogen**.
(6) In der Hoffnung **ausländische Investoren anzuziehen** halten lateinamerikanische Regierungen hartnäckig daran fest, …
(7) Er **zog den Stuhl** näher **an das Bett** …
(8) **Ich duschte** und **zog** mehrere Schichten **Kleider an**.
(9) Das Nachthemd (…) war zum braungebrannten Bauch gerutscht, weil sie die **Knie** bis fast zum Kinn **angezogen hatte**.

Die Kollokationsanalyse sortiert nach logDice bezeugt auch die hohe Frequenz einiger verbo-nominalen Verbindungen und damit auch die Wichtigkeit, diese im Unterricht zu vermitteln:

(10) Noch **ein** weiterer ermutigender **Aspekt ist in Betracht zu ziehen**, …
(11) Daraus **muss der** traurige **Schluss gezogen werden**, dass dem mangelnden Interesse an Fragen der Gleichstellung der Geschlechter …

Die Konkordanzen erscheinen auch mit parallelen Passagen in einer anderen Sprache, falls diese am Anfang der Recherche ausgewählt wurde.

Fazit:
Aus beiden Recherchen ist ersichtlich, in welchen Verbindungen das Verb *anziehen* in deutschen geschriebenen Texten vorkommen kann. Die signifikanten Kollokationspartner in den Abb. 133 und 134 bieten Ausgangspunkte für weiterführende Arbeiten auf der syntaktischen und lexikalischen Ebene an, die sich auf die textuelle Ebene erweitern lassen (vgl. Punkt 3 auf der folgenden Seite).

Die Recherchen zeigten:

1) syntaktisch-semantische Rollen, die sich um das Prädikat *anziehen* verbreiten:
   [*eine Person*] *zieht* ([*sich*/ etwas]) *an*
   [*etwas*] *anziehen*
   [*ein Ereignis*] *zieht* [*Personen*] *an* etc.
2) semantische Gruppen von Substantiven, mit denen das Verb *anziehen* häufig vorkommt (hier in Rolle des Patiens):
   a) *Besucher/ Massen/ Scharen/ Publikum/ Touristen anziehen*
   b) *Zügel/ Handbremse/ Schraube anziehen*
   c) *Kleider*, [*warme, bequeme*] *Kleidung anziehen*
3) Prädikatsverben, die häufig in der Umgebung des Verbs *anziehen* vorkommen (*waschen, duschen*), sind Zentren benachbarter Prädikationen. Gemeinsam mit der Prädikation mit *anziehen* gewährleisten sie die textuelle Kohärenz. Eine solche ikonographische Schilderung (*sich waschen, dann* [*etwas*] *anziehen*) ist in den Texten offensichtlich nicht selten.

Diese Schlussfolgerungen müssten noch nach Wahl der Korpora, in denen recherchiert wurde, differenziert werden. Sie zeigen aber deutlich die Möglichkeiten von Interpretationen, die aus Ergebnissen der automatisch berechneten Kookkurrenzen/ Kollokationen gezogen werden können.

Bemerkung zur Studie 10

Zum Schluss dieser Studie muss nochmals die Kookkurrenz-/ bzw. Kollokationsanalyse hervorgehoben werden. Sie stellt eine wichtige, sogar revolutionäre Funktion für die Bereiche der Erforschung der Sprache(n) in gleich mehreren Aspekten dar:

Im (Fremd-)Sprachenunterricht ist sie grundlegend für die Ermittlung von „wahrscheinlichen Wortkombinationen“ in der Sprache (Westhoff 1991: 16), derer Kenntnis das Westhoffsche (1987: 41) „dritte Redundanzfeld“ bilden.

Des Weiteren beschränken sich die Möglichkeiten der Kollokations-/ Kookkurrenzanalyse nicht nur auf die (in dieser Studie und in diesem Buch betonte) linguistische, bzw. linguodidaktische Ebene. Sie hilft unter anderem auch, die metaphorische Sprache der Börsianer (vgl. Lišková 2010: 25) aufzudecken und besser zu verstehen. Eine Kookkurrenzanalyse zu *Aktienkurs* im DeReKo deckt (auch für einen Wirtschaftslaien) interessante syntagmatische Muster auf. Hier werden nur die bildhaftesten genannt, weitere siehe Lišková 2010.

der Aktienkurs sackte (ab/ein)
dümpelt (vor sich hin)
steigt/ schnellt (in die Höhe/ nach oben)
bricht ein
beflügelt jmdn..

Durch die Berechnung der typischen Kollokationen aus Zeitschriften- und Zeitungstexten lassen sich auch politische, soziale und kulturelle Einstellungen der sprachlichen Kommunität ableiten. Ein Vergleich der signifikanten Kollokationspartner zum Wort *anpassungsfähig* (*anpassungsunwillig*) in drei Korpora mit Zeitungstexten in drei Sprachen ist verblüffend:
Während in deutschen Texten (Subkorpus Deutscher Zeitungen und Zeitschriften aus den Jahren 2004-2013) diese Wörter relativ selten und am häufigsten mit dem Substantiv *Wesen* (*anpassungsfähiges/ anpassungsunwilliges Wesen*) verbunden sind, haben seine tschechischen

und slowakischen lexikalischen Entsprechungen (*nepřizpůsobivý*, bzw. *nepriespôsobivý*) in publizistischen Texten eine hohe Frequenz und diese signifikanten Partner:

> *sociálně, neplatič ... spoluobčan, občan ... Rom ... problémový ...ghetto*
> *neplatič, sociálne ... asociál ... bezdomovec, občan, cigán ... etnikum*

Anpassungsunfähig/ -unwillig sind in tschechischen und slowakischen publizistischen Texten also in erster Linie Menschen, sogar konkret die Roma-Minderheit.

Wenn man die Analyse „umdreht" und in den deutschen Texten das Lemma *Türke,* also die Bezeichnung der größten Minderheit in Deutschland abfragt, bekommt man syntagmatische Muster, die in der Abb. 135 zu sehen sind.

#	Total	Anzahl	LLR	Kookkurrenzen	syntagmatische Muster
3	425	255	111	**Griechen**	64% Griechen [und] Türken
4	612	187	88	**Kurden**	66% zwischen Türken [und] Kurden
6	744	117	72	**Araber**	77% Türken [und] Araber
7	1100	356	72	**lebenden**	99% Deutschland|hier lebenden [...] Türken
8	1111	11	60	**Türkischstämmigen**	100% Prozent der Türken und Türkischstämmigen
13	1246	26	49	**eingebürgerten**	100% von eingebürgerten [...] Türken
14	1270	24	48	**Türkinnen**	75% Türkinnen und Türken
15	1291	21	41	**bewohntes**	100% ein von Türken bewohntes Haus
16	1299	8	41	**niederbayrische**	100% der niederbayrische Türke mit ...
17	1324	25	37	**eingebürgerte**	100% eingebürgerte Türken
41	1589	15	14	**aufgewachsenen**	100% Deutschland|und aufgewachsenen [...] Türken
42	1619	30	13	**geborener**	90% Deutschland geborener [...] Türke und ein

Abb. 135: Kookkurrenzen und syntagmatische Muster zum Lemma *Türke* (DeReKo)

Diese Beispiele zeigen, dass Korpusdaten auch für Politologen und Soziologen eine wichtige Forschungsquelle darstellen können, denn die Sprache ist bekanntlich ein Spiegel der Gesellschaft.

Studie 11: Quasi-Anglizismen

Der Vergleich von Sprachen führt zu einer höheren Sprachbewusstheit, diese „stellt den ersten wichtigen Schritt zum eigenverantwortlichen Umgang mit Textprodukten und Lernprozessen dar“ (Sorger et al. 2013: 290), dadurch werden die Sprachkompetenzen der Lerner/-innen erhöht. Der Vergleich von mehreren Sprachen auf der lexikalischen Ebene kann kulturelle und soziale Kontakte, bzw. gegenseitige Beeinflussungen und oder Irrtümer zeigen.

Zum Vergleich von Lexemen reichen oft gängige zweisprachige Wörterbücher[54]. Nun können diese aber erstens nie alle Lexeme der Sprache beinhalten, zweitens können sie auch nicht genug Kontext liefern. Dieses Manko lässt sich durch Korpusrecherchen ausgleichen, wie diese Studie über scheinbare Anglizismen zeigt.

In den mitteleuropäischen Sprachen gibt es Modewörter, die das „modische“ Englisch bloß nachahmen, bzw. die gängigen englischen Formen (leicht) verändern. Solche Wörter bilden sog. Faux amis, weil sie im Englischen in der Form, die den Englischlernenden „vertraut“ ist, gar nicht existieren. Zu ihnen gehören im Deutschen *Aircondition* (weniger üblich *Air Condition*), *Happy End* (oder weniger üblich *Happyend*)[55], *Oldtimer*, *Showmaster* oder das allgegenwärtige und fast notorische *Handy*. Bis auf *Handy* handelt es sich tatsächlich um „Mitteleuropäismen“, denn die meisten von ihnen sind in Nationalen Korpora einiger mitteleuropäischer Sprachen belegt, wie aus der Tab. 28 ersichtlich ist. Die Schreibweise wurde im originalgetreuen Bild der jeweiligen Sprache beibehalten. In Klammern sind die absoluten Anzahlen der Belege angegeben. Sie sollen hier nur die Existenz der Formen bestätigen, sagen jedoch wenig über die Frequenz der Wörter in der jeweiligen Sprache aus.

Slowenisch FidaPLUS	**Polnisch** Nacijonalnyj korpus Jezyka Polskeigeo	**Ungarisch** Magyar Nemzeti Szövegtár	**Tschechisch** Český národní koprus	**Slowakisch** Slovenský národný korpus
happy end *happyend* (2)	*happy end* (133)	*happyend* (2)	*happy end* (1450) *happyend* (522)	*happy end* (283) *happyend* (204)
oldtimer (42)	*oldtimer* (3)	*oldtimer* (18)	*oldtimer* (322)	*oldtimer* (19)
air condition (37)	*air condition* (4)	*air kondi* (2)	*aircondition* (5)	*aircondition*
showmaster (1)	-	-	*showmaster* (13)	*show master* (2) *šoumáster* (10)

Tab. 27: Quasi-Anglizismen in einigen Korpora mitteleuropäischer Sprachen.

Die Quasi-Anglizismen *happy end* und *air condition* beschränken sich nicht nur auf die hier betrachteten Sprachen. Sie sind auch im Türkischen Nationalkorpus belegt, wie einige Konkordanzzeilen in den Abb. 136 und 137 bezeugen.

No	Text	Query Results
1	AG03A1B-3172	genellikle mutlu bir sonla noktalanır (**happy end**). Aslında herşeyin planlı,
2	HG09C2A-0593	perdenin ardından gözlerindeki yaşı silecektir… **Happy End**!… Film,
3	DG03C4A-1508	(yanlış) aşılama yöntemi, sanıldığı gibi "**HAPPY END**" ile sonuçlanmıyor.

Abb. 136: Konkordanzen zur Abfrage *happy end* (Türkçe Ulusal Derlemi)

[54] Bis auf *Aircondition* sind die hier analysierten im Pons Globalwörterbuch Deutsch-Englisch (1998) aufgelistet.

[55] Die Schreibweisen *Aircondition*/ *Air Condition*, *Happy End*/ *Happyend* wurden an Daten von DeReKo überprüft.

No	Text	Query Results
1	NG24D1B-2301	Yazları sıcaktan bunalınca **air condition** cenneti kütüphanelere takılın.
2	GD02A3A-1321	, kuru temizlemeden, elektrikli battaniye, **air condition** ve aya insan ay
3	BD02A3A-2967	… distribütörün yanından geçirip **air condition** kompresörüne bağlayın
4	MD30D1B-2199	gün toplanan MGK'daki havayı, "**Air condition** çalışıyordu, hiç üşümedik"
5	ED02A2A-1023	mutfak, tuvalet, soğukhava tertibatı "**air condition**" gibi günlük ihtiyaçları
6	MD30D1B-2199	Toplantıda hava gayet iyiydi. **Air condition** bile çalışıyordu.
7	EA16B1A-1745	bir koltuk. Dolby stereo sistem. **Air condition**. İşte camlı bölmede ve

Abb. 137: Konkordanzen zur Abfrage *air condition* (Türkçe Ulusal Derlemi)

Wie die Entsprechungen im gleichen Text, jedoch in einer anderen Sprache aussehen können, lässt sich entweder aus einem Parallelkorpus oder aus einem Instrument, das auf parallelen Texten aufgebaut ist (z.B. linguee.de), feststellen.

Fragestellung: *Was entspricht dem deutschen Wort* Oldtimer *im Englischen und im Tschechischen?*

Die Suche wird im **InterCorp** durchgeführt (im Opus-Corpus gibt es keine Belege zu *Oldtimer*.)

- **korpus.cz** → Login → KonText → Parallel corpus InterCorp → intercorp_de
- Aligned corpora → intercorp_en → Add
- Aligned corpora → intercorp_cs → Add
- Query type: Basic
- Basic (Suchfenstereingabe): **Oldtimer** → Search

Die Ergebnisse erscheinen in Form einer Tabelle mit den ausgewählten Sprachen. Die KWICs sind in der Ausgangssprache rot markiert, in den anderen Sprachen müssen sie gesucht werden. In der Abb. 138 sind sie mit Fettschrift hervorgehoben.

de	cs	en
[Er] hatte die Boxen zu Unterstellplätzen für nagelneue Nobelkarossen und wertvolle **Oldtimer** umgebaut.	Stáj byla přeměněna v impozantní garáž a zaparkované vozy tvořily opravdu sbírku prvotřídních značek: černé ferrari, (…) - a **historické Porsche** 356.	The stalls had been converted into an impressive automotive parking facility. The collection was astonishing - a black Ferrari (…), a **vintage Porsche** 356.
… zur Verwendung in älteren, besonders beschaffenen Fahrzeugen (**Oldtimer**)	[prodej] pro použití ve **starých vozidlech** zvláštních vlastností	… sales to be used by **old vehicles** of a characteristic nature
(10) **Oldtimer**, d. h. historische Fahrzeuge, Fahrzeuge mit Sammlerwert oder	(10) **Dobová vozidla**, to znamená historická vozidla nebo vozidla se sběratelskou hodnotou nebo	(10) **Vintage vehicles**, meaning historic vehicles or vehicles of value to collectors
Und selbst als **Oldtimer** für kleine Ausflüge am Sonntag taugt der Käfer von Anfang der 70er …	Brouk ze začátku 70 . let se moc nehodí ani jako **veterán** na krátké nedělní výlety …	And even for Sunday drives the Beetle of the early 1970s can hardly be regarded as a **vintage car** …
Obwohl eine kleine Gruppe von Sammlern spekulativ in **Oldtimer** oder Automobilraritäten investiert …	Třebaže malá skupina sběratelů spekulativně investuje do **veteránů** a nevšedních vozů …	Though a small group of collectors invests speculatively in **antique** or specialty **cars** …

Abb. 138: Parallele Textpassagen zur Abfrage *Oldtimer* (InterCorp de – cs – en. Hervorhebungen Autor)

Analog zu dieser Recherche erfolgt auch die Suche nach andern Quasi-Anglizismen. Es können auch mehrere Sprachen hinzugefügt werden.

Fazit:
Auch ein/e wenig erfahrene/r Schüler/in schlussfolgert aus den parallelen Passagen, dass der „mitteleuropäische Anglizismus“ *Oldtimer* im Englischen nicht existiert, sondern dass ihm die Verbindung mit dem Attribut *vintage* oder *antique* am ehesten entspricht.

Im Tschechischen findet man hier als Entsprechung zum deutschen *Oldtimer* nur die Bezeichnung *veterán*. Das Wort *oldtimer* wird im Tschechischen zwar (fast) in demselben Sinn, vielleicht aber seltener verwendet, als sein deutsches Pendant. Wie die prozentuelle Verteilung der Entsprechungen in einzelnen Texten zueinander stehen, muss aufgrund der zu dieser Recherche immer noch geringen Korpusgröße vorerst unbeantwortet bleiben.

Studie 12: Ein Blick in die Geschichte vom *Hit*

An Korpusdaten lassen sich sogar einige Änderungen in der Sprache beobachten. Als Beispiel für solche Recherchen wird die Änderung im Gebrauch der Wörter *Schlager* und *Hit* gezeigt.

Forschungsfrage: *Wann hat eigentlich der* Hit *den* Schlager *abgelöst?*

Die Antwort auf diese Frage zu Diachronie der Sprache kann natürlich niemand auf das Jahr oder sogar den Tag genau sagen. Die Änderungen im Wortschatz verlaufen auf verschiedenen Wegen: der *Hit* ist zu einem Modewort geworden, hat den *Schlager* aus seiner Position geschlagen, aber der *Schlager* konnte sich behaupten und mit einer leicht veränderten Bedeutung ist er auch im 21. Jh. ein fester Bestandteil der deutschen Sprache (vgl. Abb. 143 auf Seite 182).
Einen groben Überblick in die Geschichte der Verwendung einzelner Wörter verschaffen am besten die Korpora **DWDS** und **DeReKo**.

Recherche im DWDS

- **DWDS** → Abfragefenster
- Suchfeldeingabe: **Hit** → Suche im DWDS
- Ressourcen → Wortverlauf (Basis DWDS-Kernkorpus) (Ergebnis: Abb. 139)

dann

- Suchfeldeingabe: **Schlager** →
- Ressourcen → Wortverlauf (Basis DWDS-Kernkorpus) (Ergebnis: Abb. 140)

Zu beiden Abfragen liefert das DWDS eine graphische Übersicht über die Streuung der Wörter in Dekaden über das 20. Jahrhundert.

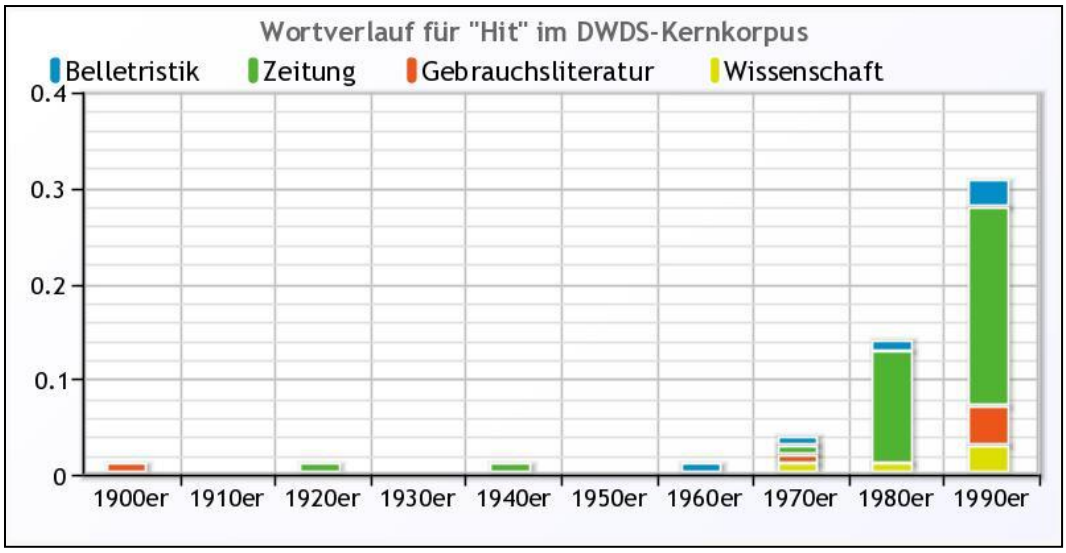

Abb. 139: Graphik: Wortverlauf von *Hit* im 20. Jh. (DWDS)

Die Belege in den 0-er, 20-er und 40-er Jahren sind Tippfehler (jeweils ein Beleg), die das Korpus falsch erkannt hat.

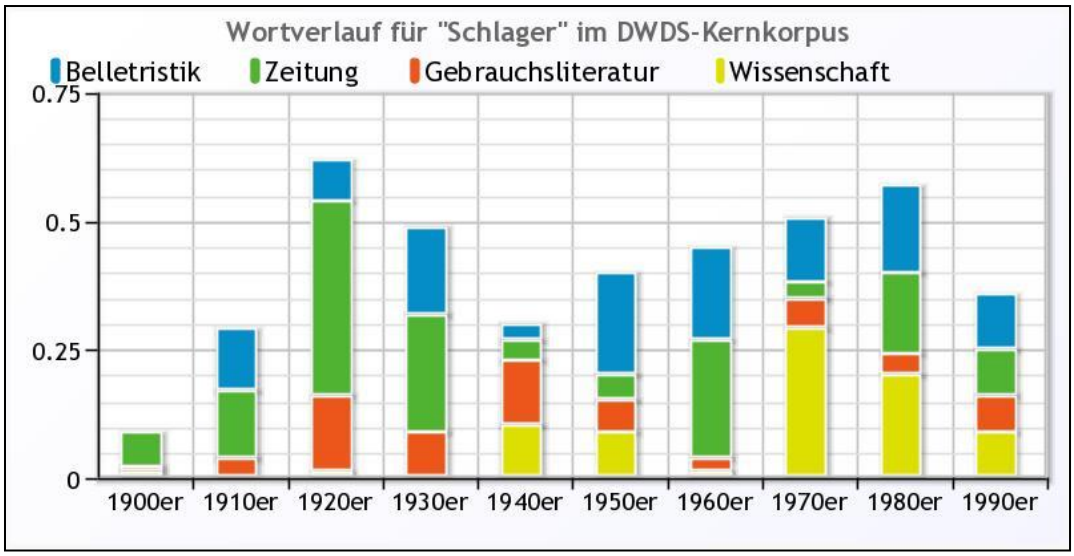

Abb. 140: Graphik: Wortverlauf von *Schlager* im 20. Jh. (DWDS)

Die Streuung des Wortes *Schlager* im 20. Jh. erinnert an eine Sinuskurve. Bei der Betrachtung der Belege stellt man Folgendes fest: In den ersten Jahrzehnten war mit *Schlager* allgemein ein künstlerisches Erfolgswerk gemeint. Dies lässt sich aus den ausgewählten Textpassagen in der Abb. 141 ableiten. In Klammern sind die Erscheinungsjahre der Texte angegeben.

(1903)	Eine "Elf Scharfrichter" - Nummer mit einem Titelbild von Th. Th. Heine ist der neueste **Schlager** , mit dem der Verlag Bühne und Brettl (…), soeben auf dem Plane erscheint.
(1914)	Um diesen **Schlager** wird die Operette gebaut, wölbt sie sich, wie die Glasglocke um den Kuhkäse.
(1922)	Dem Schauspieler wurde es immer flauer ums Herz. Vor diesen Leuten sollte er sich nun produzieren, sollte lustige **Schlager** zum besten geben. In einem Kaffeehaus…
(1922)	Der **Schlager** des Tages aber sollte die Ausstellung des damals eben gezüchteten Rassepferdes "Westfälisches Edelblut" sein.

Abb. 141: Auswahl der Konkordanzen zur Abfrage *Schlager* (DWDS, Kernkorpus)

Die Abfragen werden zum Vergleich im DeReKo durchgeführt.

Recherche im DeReKo

Der Korpusmanager Cosmas II ermöglicht nämlich auch die Jahrzehntansicht, darüber hinaus aber auch die Ansicht nach Jahren (Jahresansicht), Monaten (Monatsansicht), sogar Tagen (Tagesansicht).

- **Cosmas II$_{web}$** → Recherche
- Archiv: W-Archiv der geschriebenen Sprache
- Korpus: W-öffentlich - alle öffentlichen Korpora des Archivs W
- **Optionen**: **Suchmodalitäten:**
  Groß- / Kleinschreibung beachten für 1. Zeichen — auswählen
  Expansionslisten: — abwählen
  Übernehmen
  Ergebnispräsentation: Jahrzehntansicht — auswählen
- **Suchanfrage** (Eingabe ins Suchfeld)**: Hit** → Suchen (Ergebnis: Abb. 142)
  dann: **Schlager** → Suchen (Ergebnis: Abb. 143)

Distribution der Wörter *Hit* und *Schlager* in Texten der 2. Hälfte des 20. Jahrhunderts

	Treffer	rel. Häuf.	Texte	Jahrzehnt
⊞	4	1.41 pMW	4	1960-1969
⊞	1	0.51 pMW	1	1970-1979
⊞	28	2.36 pMW	27	1980-1989
⊞	9.915	10.19 pMW	9.105	1990-1999
⊞	13.855	7.75 pMW	12.564	2000-2009
⊞	15.449	10.33 pMW	12.224	2010-2019

Abb. 142: Wortverlauf von *Hit* (DeReKo)

	Treffer	rel. Häuf.	Texte	Jahrzehnt
⊞	3	6.89 pMW	3	1940-1949
⊞	5	2.82 pMW	3	1950-1959
⊞	28	9.85 pMW	21	1960-1969
⊞	14	7.09 pMW	13	1970-1979
⊞	39	3.29 pMW	35	1980-1989
⊞	15.122	15.54 pMW	12.096	1990-1999
⊞	16.047	8.97 pMW	12.919	2000-2009
⊞	10.431	6.97 pMW	8.064	2010-2019
	41.689	**9.74 pMW**	**33.154**	**8 Jahrzehnte**

Abb. 143: Wortverlauf von *Schlager* (DeReKo)

Zur Ansicht nach Jahren wechselt man durch Klicken auf Ergebnisse → Jahresansicht am linken Rand.

Fazit

Die Daten in den Abb. 142 und 143 bestätigen die grobe Darstellung der Abb. 139 und 140: es ist deutlich zu sehen, dass der *Hit* etwa Anfang der 1970er Jahre den *Schlager* abgelöst hat. (Konkordanzen aus den einzelnen Jahrzehnten werden durch einen Klick auf ⊞ in der entsprechenden Zeile abgerufen.) Interessant wäre noch der Vergleich von Kookkurrenzen von *Schlager* in den 1940ern bis 1960ern und *Hit* ab den 1970er Jahren. Diese Untersuchung bleibt vorerst ein Desiderat.

Pragmatik

Mithilfe von Korpora lassen sich relativ gut auch pragmatische Elemente der Sprache abrufen, solange diese Elemente irgendwie fest an der sprachlichen Oberfläche verankert sind.

Studie 13: Illokutionsverben

Zur Manifestation eines pragmatischen Markers zählen Prädikate in Einleitungssätzen zu direkter oder indirekten Rede. Diese können nur in morphosyntaktisch annotierten Korpora recherchiert werden (hier **DeReKo**), indem das Prädikatsverb abgerufen wird. Die Fragestellung nach den Realisierungsmöglichkeiten dieser Verben eröffnet auch die Hypothese, dass diese Verben nicht alle dem Wortfeld/ der semantischen Gruppe der Verben des Sprechens (verba dicendi) angehören müssen (vgl. Hirschová 2013: 172-174).

- **Cosmas II$_{web}$** → Recherche
- Archiv: TAGGED-T - Archiv morphosyntakt. annotierter Korpora (TreeTagger)
- Korpus: TAGGED-T-öffentlich - alle öffentlichen Korpora des Archivs TAGGED-T
- **Optionen**: Standardeinstellungen
- **Suchanfrage** (Eingabe ins Suchfeld)**: MORPH(VRB·fin·v)·:·\"** → Suchen

dann:

- **Suchanfrage** (Eingabe ins Suchfeld)**:**
  MORPH(VRB·fin·v)·,·/+w1·dass·oder·daß → Suchen

Die erste Abfrage bedeutet: Suche alle finiten Verben gefolgt von einem Doppelpunkt und Anführungszeichen. Das Ergebnis (Auswahl) ist in der Abb. 144 zu sehen.
Die zweite Abfrage bedeutet: Suche alle finiten Verben gefolgt von einem Beistrich und *dass*, bzw. *daß*. Einige Ergebnisse dieser Abfrage sind in der Abb. 145 zu sehen.

Die Einleitungen bzw. „Ausleitungen" der direkten Rede werden u.a. durch diese Prädikate realisiert:

HMP09	Und auch Sturm-Partner Klose	**bemerkte erleichtert: "Ich**	bin froh,
NON09	Truppe gibt. An die Kritiker	**appelliert er: "Wir**	befinden uns mit
RHZ09	zu führen. Der CDU-Politiker	**betonte zugleich: "Die**	Straße ist fast
NON08	und Ziele für Peking	**sagt er: "Ich**	erwarte nicht, denn
RHZ08	zum Koblenzer Stadtrand, dann	**sagt sie: "Auf**	geht's. Du läufst heute
M07	der Radsport Doping-verseucht?"	**antwortete Ullrich: "Das**	wüsste ich
NON07	Leopold Schmölz	**ist zuversichtlich: "Mit**	der Gemeinde
RHZ07	erklärt Klute-Wetterauer und	**fügt hinzu: "Wir**	wollen dem Nachwuchs
HMP06	Corinna Berghoff - selbst unverheiratet -	**lächelt vieldeutig: "Darauf**	warte ich
RHZ06	Elsbeth Schmidt	**zog Erfolgsbilanz: "Wir**	haben den
RHZ06	herum. Trainer Burkhard Lau	**meint aber: "Abgerechnet**	wird am
RHZ06	Trainer Stefan Liesenfeld	**ärgerte sich: "Von**	den Chancen her

Abb. 144: Auswahl der Konkordanzen zur Abfrage: finites Verb – Doppelpunkt - Anführungszeichen (DeReKo)

Diese Prädikate können mit jenen verglichen, die die indirekte Rede einleiten.
Die indirekte Rede kann im Deutschen durch die Prädikate in der Abb. 145 (nächste Seite) ein- oder ausgeleitet werden.

NUN9	Die Autoren	**bemerken kritisch, dass** der (westlichen) melancholischen Sehnsucht
MO8	…alle (…)	**sagen** uns**, dass** wir viel zu streng sind.
A09	Gemeindepräsident Hans Bütikofer	**hielt fest, dass** er und seine Amtskollegen die
A09	Remer und seine Kollegen sagen (…)und	**warnen** zugleich, **dass** in bestimmten Regionen -
A09	Konkordanzregierung mitzumachen. Ich	**denke nicht, dass** die SVP irgendjemandem, schon
A09	Militärsprecher behauptet habe. Weiter	**erzählte Sanur, dass** er Material und Werkzeug
A09	und Maximalsteuerfüsse vorschreibe. Zudem	**bemängelten sie, dass** die Regelung lediglich
A09	die uns gar nicht passen», sagte Grau,	**betonte aber, dass** Veränderungen zu jedem Leben
A09	Gemeindepräsident Hans Bütikofer	**hielt fest, dass** er und seine Amtskollegen die
A09	nicht als Frühaufsteherin bezeichnet,	**erklärte sie, dass** man einmal im Jahr eine
A09	muss aufrechterhalten werden.» Daneben	**betonte er, dass** das frühe Aufstehen für ihn

Abb. 145: Auswahl der Konkordanzen zur Abfrage: finites Verb – Doppelpunkt - *dass/daß* (DeReKo)

Die Abfragen in Cosmas II liefern über 62.000 Treffer für die direkte Rede (eine Auswahl davon ist in der Abb. 144 abgedruckt) und insgesamt über eine halbe Mio. Treffer für die indirekte Rede (Auswahl in der Abb. 145). Unter ihnen sind auch etliche Belege, die der gewünschten Abfrage nicht entsprechen, etwa:

*Sein Ziel **ist** es, **dass** es künftig weniger*
*für Gebäude im Mittelpunkt **stehen werde**: "Was bedeutet der*

Das Aussortieren solcher Fehlbelege dauert jedoch wesentlich kürzer als eine „manuelle“ Recherche mittels Durchlesen von Büchern, Zeitschriften, Zeitungen und anderen Texten.

Fazit:
Die Prädikate der einleitenden Sätze entsprechen der Intention der Aussage, die sie einleiten (sie entsprechen der illokutiven Kraft der Aussage). Sie spiegeln gewissermaßen die Einstellung der Sprechenden zum Sachverhalt der Aussage wider. Diese Einstellung ist aus den Prädikaten der Einleitungssätze ersichtlich. In der Tab. 29 werden sie alle aus den Abb. 144 und 145 zusammengefasst und alphabetisch aufgelistet

Prädikat	**Illokution**
Ullrich antwortete	ANTWORTEN
er appelliert	APPELLIEREN
Trainer ärgerte sich	(sich) ärgern (= SAGEN VERÄRGERT)
sie bemängelten	BEMÄNGELN
Partner, Autoren bemerkte(n) (erleichtert, kritisch)	BEMERKEN
(Politiker) betonte	BETONEN
denkt (nicht)	DENKEN/ NICHT DENKEN
Schmidt zog Erfolgsbilanz	Erfolgsbilanz ziehen = ERKLÄREN STOLZ, PRAHLEN
sie erklärt	ERKLÄREN
Sanur erzählt	ERZÄHLEN
Gemeindepräsident hielt fest	FESTHALTEN
Klutte-W. fügt hinzu	HINZUFÜGEN
Corinna Berghoff lächelt	lächeln = (Implikatur) SAGEN FRÖHLICH
Trainer meint	MEINEN
sie sagt	SAGEN
Remer und seine Kollegen warnen	WARNEN
L. Schmölz ist zuversichtlich.	zuversichtlich sein = SAGEN ZUVERSICHTLICH, OPTIMISTISCH

Tab. 29: Auswahl der Prädikate in den Einleitungssätzen zur direkten und indirekten Rede

Vergleicht man die Prädikate, die die direkte und indirekte Rede einleiten, stellt man fest, dass sie sich oft entsprechen: *bemerken, betonen, sagen* um nur einige, die für die Abb. 144 und 145 ausgewählt worden sind, zu nennen.

Es gibt natürlich weit mehr Möglichkeiten, die Intention der Sprechenden im Einleitungssatz darzustellen. Wie diese in Rundfunknachrichten realisiert werden, beschreibt Káňa (2007). Für die Analyse der Nachrichten musste ein eigenes Korpus erstellt werden, da in keinem elektronischen Korpus Radiokurznachrichten vorhanden sind. Für den allgemeinen Sprachgebrauch im DaF/DaZ-Unterricht sind die üblichen Korpora jedoch völlig ausreichend und die Lernenden können die Sprache aus diesem pragmatischen Gesichtspunkt erfassen.

7. Statistiken

Statistische Angaben über eine Sprache umreißen die Konturen dieser Sprache. Aus ihnen können wichtige Elemente für die Vermittlung der Sprache abgeleitet werden. Bisherige Werke, die sich in irgendeiner Hinsicht mit Häufigkeiten befassten (Wörterbücher, (Lerner)Grammatiken, Frequenzwörterbücher), sind veraltet und „viele davon wenig repräsentativ für die deutsche Sprache in ihrer Gesamtheit im gesamten deutschsprachigen Raum“ (Tschirner: 2005: 136). Die folgenden Angaben betreffen nur einige wenige Aspekte des Deutschen, beschränken sich mehrheitlich auf die Lexik und (teilweise) ihre Morphologie. Andere Aspekte müssen vorerst ein Desiderat bleiben, mit der Hoffnung, dass sie auch bald statistisch erfasst werden (können).

Die Angaben über die deutsche Sprache in den kommenden Tabellen basieren auf Daten des Korpus **InterCorp_de** (Gesamtkorpus) sowie auf Daten des **deTenTen-Korpus.** (Nur im letzteren lassen sich einige Kategorien sinnvoll abfragen.) Die Ergebnisse wurden teilweise (falls es die Eigenschaften des COSMAS II ermöglichen) mit Rechercheergebnissen aus **DeReKo** verglichen. So sollen die Daten objektiviert werden und einen Leitfaden anbieten, welche Elemente im Deutschunterricht auf keinen Fall fehlen sollten.

Die Tabellen sind entweder nach der Häufigkeit oder alphabetisch absteigend gereiht. Die Frequenzangaben beziehen sich auf die relative Häufigkeit umgerechnet auf 1 Mio. Wörter (in der Sprache des InterCorp „i.p.m. – instances per million“, im DeReKo „p.M.W. - pro Million Worte“).

7.1 Morphologische Kategorien

Genus der Substantive

Feminina	76 562 p.M.W.
Maskulina	64 479 p.M.W.
Neutra	42 428 p.M.W.

Kasus der Substantive (alle Genera)

Nominativ	59 542 p.M.W.
Dativ	53 000 p.M.W.
Akkusativ	51 913 p.M.W.
Genitiv	18 900 p.M.W.

Numerus der Substantive (alle Genera)

Singular	131 153 p.M.W.
Plural	52 718 p.M.W.

Die häufigste morphologische Form der Substantive ist **Nominativ Singular Maskulinum**, gefolgt von **Nominativ Singular Femininum**.
Die vergleichsmäßig „seltensten“ Formen sind Genitive Plural (feminin, maskulin, neutrum).

Komparationsformen der Adjektive

Positiv	46 559 p.M.W.
Komparativ	1 523 p.M.W.
Superlativ	1 522 p.M.W.

Artikel

Bestimmte Artikel	70 016 p.M.W.
Unbestimmte Artikel	17 049 p.M.W.

Frequenzen der Verben nach ihrer Form und Funktion:

Finite Vollverben	etwa 43%
Finite Hilfsverben	etwa 22%
Partizip Perfekt von Vollverben	etwa 11%
Infinite Vollverben	etwa 11%
Finite Modalverben	etwa 6%
Infinite Hilfsverben	etwa 2%
(Fehler	etwa 5%)

7.2 Lexikalische Realisierungen

Substantive mit der höchsten Frequenz (alphabetisch):

1. Abend
2. Anfang
3. Arbeit
4. Art
5. Arzt
6. Aufgabe
7. Auge
8. Auto
9. Beispiel
10. Bild
11. Blick
12. Buch
13. Deutsch/e
14. Eltern
15. Ende
16. Fall
17. Familie
18. Form
19. Frage
20. Frau
21. Freund
22. Geld
23. Gesellschaft
24. Gesetz
25. Geschichte
26. Gesicht
27. Gespräch
28. Grenze
29. Grund
30. Gruppe
31. Hand
32. Haus
33. Herr
34. Herz
35. Hilfe
36. Idee
37. Jahr
38. Jahrhundert
39. Junge
40. Kampf
41. Kind
42. Kirche
43. Kopf
44. Kraft
45. Krieg
46. Land
47. Leben
48. Leute
49. Mädchen
50. Mal
51. Mann
52. Mensch
53. Meter
54. Minute
55. Mitte
56. Möglichkeit
57. Monat
58. Mutter
59. Nacht
60. Name
61. Ort
62. Partei
63. Person
64. Platz
65. Problem
66. Punkt
67. Raum
68. Recht
69. Richtung
70. Sache
71. Seite
72. Schritt
73. Schule
74. Sinn
75. Sohn
76. Spiel
77. Staat
78. Stadt
79. Stelle
80. Stimme
81. Straße
82. Stück
83. Stunde
84. Tag
85. Teil
86. Tier
87. Tod
88. Tochter
89. Uhr
90. Vater
91. Wasser
92. Weg
93. Welt
94. Wohnung
95. Woche
96. Wort
97. Zeit
98. Zeitung
99. Ziel
100. Zug

Adjektive mit der höchsten Frequenz (alphabetisch):

1. absolut
2. ähnlich
3. allgemein
4. allmählich
5. alt
6. angeblich
7. angenehm
8. ängstlich
9. anscheinend
10. anständig
11. arm
12. aufmerksam
13. ausgerechnet
14. bekannt
15. bequem
16. bereit
17. berühmt
18. bescheiden
19. besorgt
20. bestimmt
21. bewuss/ßt
22. billig
23. bitter
24. blass/ß
25. blau
26. bleich
27. blöd
28. blond
29. bloß
30. böse
31. braun
32. brav
33. breit
34. bunt
35. dankbar
36. dauernd
37. deutlich
38. deutsch
39. dick
40. dicht
41. direkt
42. doppelt
43. dringend
44. dritt
45. dumm
46. dumpf
47. dunkel
48. dünn
49. düster
50. egal
51. ehrlich
52. echt
53. eifrig
54. eigen
55. eilig
56. eindeutig
57. einfach
58. einsam
59. einzeln
60. einzig
61. elegant
62. elend
63. endgültig
64. endlos
65. eng
66. englisch
67. entfernt
68. entsetzlich
69. entscheidend
70. entsprechend
71. erfolgreich
72. genau
73. gleich
74. groß
75. gut
76. hoch
77. jung
78. kalt
79. klar
80. klein
81. kurz
82. lang
83. leer
84. leicht
85. letzt
86. möglich
87. nah
88. neu
89. offen
90. öffentlich
91. politisch
92. richtig
93. rund
94. schnell
95. schön
96. schwer
97. stark
98. vergangen
99. weiß
100. wichtig

Vollverben mit der höchsten Frequenz (alphabetisch)

1. ändern
2. arbeiten
3. aufnehmen
4. aussehen
5. bauen
6. befinden
7. beginnen
8. bekommen
9. berichten
10. bestehen
11. bieten
12. bleiben
13. brauchen
14. bringen
15. denken
16. entscheiden
17. entstehen
18. erfahren
19. erhalten
20. erinnern
21. erklären
22. erreichen
23. erscheinen
24. erwarten
25. erzählen
26. fahren
27. fallen
28. fehlen
29. finden
30. folgen
31. fordern
32. fragen
33. freuen
34. fühlen
35. führen
36. geben
37. gehen
38. gehören
39. gelten
40. geraten
41. gewinnen
42. glauben
43. halten
44. handeln
45. heißen
46. helfen
47. hoffen
48. holen
49. hören
50. kennen
51. kommen
52. lassen
53. laufen
54. leben
55. legen
56. leisten
57. lernen
58. lesen
59. liegen
60. machen
61. meinen
62. melden
63. nehmen
64. nennen
65. öffnen
66. reden
67. rechnen
68. sagen
69. sehen
70. setzen
71. schaffen
72. scheinen
73. schlagen
74. schließen
75. schreiben
76. sitzen
77. sorgen
78. spielen
79. sprechen
80. stehen
81. steigen
82. stellen
83. sterben
84. suchen
85. tragen
86. treffen
87. tun
88. übernehmen
89. verlassen
90. verlieren
91. verstehen
92. versuchen
93. vorstellen
94. wählen
95. warten
96. wirken
97. wissen
98. zählen
99. zeigen
100. 100. ziehen

Die Verben *haben* und *sein* als Vollverben befinden sich erst unter den ersten 120 häufigsten deutschen Verben.

Modalverben

(Anteil am Gesamtvorkommen der Modalverben)

1. können ……..… 37 %
2. müssen ……..…… 25 %
3. sollen ..…… 18 %
4. wollen ..……..…… 8 %
5. mögen …… 7 %
6. dürfen …… 5 %

100%

Partizipien Perfekt mit der höchsten Frequenz (nach Frequenz):

1. gesagt
2. gesehen
3. gekommen
4. gemacht
5. gegeben
6. getan
7. gefunden
8. gebracht
9. gegangen
10. verloren
11. gedacht
12. gesprochen
13. genommen
14. vergessen
15. geschrieben
16. verlassen
17. geblieben
18. gehört
19. verschwunden
20. gefallen
21. erfahren
22. geschehen
23. passiert
24. gehalten
25. genannt
26. gefragt
27. erreicht
28. gelesen
29. geboren
30. versucht
31. gestellt
32. gelernt
33. gestorben
34. geführt
35. verstanden
36. erwartet
37. gesetzt
38. gelegt
39. getroffen
40. gezogen
41. entdeckt
42. gewuss/ßt
43. geschlossen
44. erklärt
45. begonnen
46. verrückt
47. beschäftigt?
48. geschlagen
49. erschienen
50. erzählt
51. gelungen
52. entfernt
53. bestimmt
54. gefahren
55. geraten
56. bemerkt
57. gerichtet
58. geglaubt
59. gezeigt
60. gewöhnt
61. erhalten
62. angefangen
63. gelassen
64. geöffnet
65. aufgenommen
66. gestanden
67. gezwungen
68. geschickt
69. gekauft
70. erfüllt
71. geworfen
72. verändert
73. erkannt
74. gelebt
75. erschrocken
76. überrascht
77. entschlossen
78. gelegen
79. versteckt
80. angenommen
81. gemeint
82. begriffen
83. getragen
84. gespielt
85. geschafft
86. gewartet
87. geliebt
88. gewonnen
89. verbunden
90. gerufen
91. bekommen
92. verboten
93. geschlafen
94. gebaut
95. versprochen
96. bekannt
97. gesucht
98. gebeten
99. vorbereitet
100. gerettet

Infinitive zu diesen häufigsten Partizipien der deutschen Verben sind auf der folgenden Seite alphabetisch aufgelistet.

Infinitive zu den häufigsten Partizipien (alphabetisch):

1. anfangen
2. annehmen
3. aufnehmen
4. bauen
5. beginnen
6. begreifen
7. bekennen?
8. bekommen
9. bemerken
10. beschäftigen?
11. bestimmen?
12. bitten
13. bleiben
14. bringen
15. ge-|denken
16. entdecken
17. entfernen?
18. entschließen
19. erfahren
20. erfüllen
21. erhalten
22. erkennen
23. erklären
24. erreichen
25. erscheinen
26. erschrecken
27. erwarten
28. erzählen
29. fahren
30. fallen|gefallen
31. finden
32. fragen
33. führen
34. gebären?
35. geben
36. gehen
37. ge-|hören
38. gelingen?
39. geraten|raten
40. geschehen
41. ge-|stehen
42. gewinnen
43. gewöhnen
44. glauben
45. halten
46. kaufen
47. kommen
48. lassen?
49. leben
50. legen
51. lernen
52. lesen
53. lieben?
54. liegen
55. machen
56. meinen
57. nehmen
58. nennen?
59. öffnen?
60. passieren
61. retten?
62. richten?
63. rufen
64. sagen
65. sehen
66. setzen
67. schaffen?
68. schicken
69. schlafen
70. schlagen
71. schließen?
72. schreiben
73. spielen
74. sprechen
75. stellen
76. sterben
77. suchen
78. tragen
79. treffen
80. tun
81. überraschen?
82. verändern
83. verbieten
84. verbinden?
85. vergessen
86. verlassen
87. verlieren?
88. verrücken?
89. verschwinden
90. versprechen
91. verstecken?
92. verstehen
93. versuchen
94. vorbereiten
95. warten
96. werfen
97. wissen
98. zeigen
99. ziehen
100. zwingen

Bemerkung:
Bei den meisten deutschen Verben gilt Folgendes: das Vorkommen vom Infinitiv zusammengerechnet mit allen finiten Formen desselben Verbs hat eine höhere Frequenz als sein Partizip Perfekt (inkl. fehlerhaft annotierte deverbative Adjektive). Umgekehrt ist es bei den Verben, die mit einem Fragezeichen gekennzeichnet sind.

Verbzusätze mit der höchsten Frequenz

Nach Frequenz

1. an
2. auf
3. aus
4. zu
5. ein
6. ab
7. zurück
8. vor
9. hin
10. um
11. zusammen
12. nach
13. weiter
14. heraus
15. fort
16. mit
17. herum
18. da
19. her
20. fest
21. hinaus
22. hervor
23. vorbei
24. hinein
25. los
26. hoch
27. hinzu
28. hinunter
29. hinauf
30. weg
31. entgegen
32. durch
33. heran
34. nieder
35. nahe
36. herein
37. gegenüber
38. hinüber
39. herunter
40. herab
41. über
42. hinab
43. bei
44. empor
45. voraus
46. inne
47. raus
48. umher
49. statt
50. auseinander

Alphabetisch

1. ab
2. an
3. auf
4. aus
5. auseinander
6. bei
7. da
8. durch
9. ein
10. empor
11. entgegen
12. fest
13. fort
14. gegenüber
15. her
16. herab
17. heran
18. heraus
19. herein
20. herum
21. herunter
22. hervor
23. hin
24. hinab
25. hinauf
26. hinaus
27. hinein
28. hinüber
29. hinunter
30. hinzu
31. hoch
32. inne
33. los
34. mit
35. nahe
36. nach
37. nieder
38. raus
39. statt
40. über
41. um
42. umher
43. vor
44. voraus
45. vorbei
46. weg
47. weiter
48. zu
49. zurück
50. zusammen

Verben mit den häufigsten Verbzusätzen

Die folgenden alphabetischen Listen der häufigsten Verben mit einzelnen Verbzusätzen wurden als ein Querschnitt von komplexen Abfrageergebnissen (siehe Studie 7) im DeReKo und InterCorp_de zusammengestellt. Die Kriterien für die Auswahl der Lexeme waren signifikante Frequenzen in beiden Korpora und bei allen Abfragen. Die Vorgangsweise bei der Recherche ermöglicht auch den Vergleich, wie produktiv einzelne Verbzusätze sind. Aus diesem Grund variiert die Anzahl der Lexeme mit einzelnen Zusätzen. Die Listen wurden für die häufigsten zehn Verbzusätze zusammengestellt, des Weiteren auch für *hinüber*- und *inne*- um die Schwankungen in der Produktivität zu verdeutlichen.

Die häufigsten Verben mit dem Zusatz *an* (alphabetisch):

anbieten
anblicken
anfahren
anfangen
angehen
angreifen
anhalten
anheben
anhören
ankommen
anlächeln
annehmen
anrufen
ansagen
ansehen
ansetzen
anschauen
anschließen
anschreien
ansprechen
anstarren
anstoßen
anziehen
anzünden

Die häufigsten Verben mit dem Zusatz *auf* (alphabetisch):

aufatmen
aufblicken
aufbrechen
auffallen
auffangen
auffliegen
auffordern
aufgehen
aufhalten
aufheben
aufhören
auflachen
auflegen
aufleuchten
auflösen
aufmachen
aufnehmen
aufpassen
aufreißen
aufrichten
aufsehen
aufsetzen
aufschlagen
aufschließen
aufschreien
aufspringen
aufstehen
aufsteigen
aufstoßen
auftauchen
aufwachen

Die häufigsten Verben mit dem Zusatz *aus* (alphabetisch):

ausbrechen
ausbreiten
ausdrücken
ausgehen
aushalten
ausholen
auskennen
auslachen
auslösen
ausmachen
auspacken
ausreichen
ausrufen
ausruhen
aussehen
ausschalten
aussprechen
ausspucken
aussteigen
ausstoßen
ausstrahlen
ausstrecken
aussuchen
austauschen
austrinken
ausweichen
ausziehen

Die häufigsten Verben mit dem Zusatz *zu* (alphabetisch):

zudecken
zudrehen
zuflüstern
zugehen
zuhören
zuklappen
zukommen
zulächeln
zulassen
zulaufen
zunehmen
zunicken
zurennen
zurufen
zusehen
zuschauen
zuschieben
zuschlagen
zuschreien
zusteuern
zustimmen
zutrauen
zuwenden
zuwerfen
zuwinken
zuzwinkern

Die häufigsten Verben mit dem Zusatz *ein* (alphabetisch):

einatmen
einbiegen
einbilden
eindringen
einfallen
eingehen
eingestehen
eingießen
eingreifen
einholen
einhüllen
einladen
einleuchten
einmischen
einnehmen
einräumen
einreden
einrichten
einsammeln
einsaugen
einsehen
einsetzen
einschalten
einschenken
einschlafen
einschlagen
einschließen
einsperren
einstecken
einsteigen
einstellen
eintreffen
eintreten
einwenden
einwerfen
einziehen

Die häufigsten Verben mit dem Zusatz *ab* (alphabetisch):

abbiegen
abbrechen
abdrücken
abfahren
abfallen
abhängen
abhauen
abheben
abholen
ablaufen
ablegen
ablehnen
ablösen
abnehmen
abreisen
abreißen
absetzen
abschalten
abschließen
abschneiden
abschütteln
abspielen
abspringen
absteigen
abstoßen
abstreifen
absuchen
abtasten
abtrocknen
abwarten
abwehren
abwenden
abwinken
abwischen
abzeichnen

Die häufigsten Verben mit dem Zusatz *zurück* (alphabetisch):

zurückbleiben
zurückblicken
zurückfahren
zurückfallen
zurückgeben
zurückgehen
zurückhalten
zurückkehren
zurückkommen
zurücklassen
zurücklaufen
zurücklegen
zurücklehnen
zurückrennen
zurückrufen
zurückschieben
zurückschreien
zurückspringen
zurückstoßen
zurücktreten
zurückweisen
zurückwerfen
zurückwinken
zurückziehen

Die häufigsten Verben mit dem Zusatz *vor* (alphabetisch):

vorbereiten
vorbeugen
vordringen
vorfahren
vorfinden
vorgehen
vorkommen
vorlegen
vorlesen
vorliegen
vornehmen
vorrücken
vorsehen
vorschieben
vorschlagen
vorstellen
vorstrecken
vortäuschen
vortreten
vorwerfen
vorziehen

Die häufigsten Verben mit dem Zusatz *hin* (alphabetisch):

hinbewegen
hinblicken
hindeuten
hindrehen
hinfahren
hinfliegen
hingehen
hinhalten
hinhocken
hinknien
hinkriegen
hinlaufen
hinlegen
hinmurmeln
hinnehmen
hinrennen
hinrutschen
hinsehen
hinsetzen
hinschauen
hinschieben
hinschwanken
hinsingen
hinstarren
hinstrecken
hinsummen
hinwandern
hinweisen
hinwerfen
hinziehen

Die häufigsten Verben mit dem Zusatz *um* (alphabetisch):

umbinden
umblättern
umblicken
umbringen
umdrehen
umfallen
umgehen
…
umhauen
umkehren
umkippen
umreißen
umsehen
umschauen
umschlagen
umschnallen
umsteigen
umstoßen
umwenden
umwerfen
umziehen

Die häufigsten Verben mit dem Zusatz *hinüber* (alphabetisch):

hinüberbeugen
hinüberblicken
hinüberfahren
hinübergehen
hinüberkommen
hinüberlaufen
hinübernicken
hinüberrufen
hinübersehen
hinüberschauen
hinüberschieben
hinüberschielen
hinüberspringen
hinüberstarren
hinüberstehen
hinübertragen
hinüberwerfen
hinüberwinken
hinüberzeigen
hinüberziehen

Die häufigsten Verben mit dem Zusatz *inne* (alphabetisch):

innehaben
innehalten

Verben, die die meisten Zusätze (aus den 10 häufigsten: *an, auf, aus, zu, ein, ab, zurück, vor, hin, um*) **an sich binden** sind auf der folgenden Seite aufgelistet.
(Wegen der Übersichtlichkeit werden die Nachbarverben abwechselnd in Fett- und Normalschrift gesetzt.)

Hochfrequentierte Verben mit den häufigsten zehn Verbzusätzen:

zurück	**blicken**
um	**blicken**
hin	**blicken**
hinüber	**blicken**
auf	**blicken**
an	**blicken**
ab	brechen
aus	brechen
auf	brechen
um	**drehen**
hin	**drehen**
zu	**drehen**
zurück	fahren
vor	fahren
hin	fahren
ab	fahren
hinüber	fahren
an	**fahren**
zurück	**fallen**
um	**fallen**
ein	**fallen**
ab	**fallen**
auf	**fallen**
zurück	gehen
vor	gehen
um	gehen
hin	gehen
ein	gehen
hinüber	gehen
zu	gehen
aus	gehen
auf	gehen
an	gehen
zurück	**halten**
hin	**halten**
aus	**halten**
auf	**halten**
an	**halten**
ab	heben
auf	heben
an	heben

ein	**holen**
ab	**holen**
aus	**holen**
zu	hören
auf	hören
an	hören
zurück	**kommen**
vor	**kommen**
hinüber	**kommen**
zu	**kommen**
an	**kommen**
hinüber	laufen
zurück	laufen
hin	laufen
ab	laufen
zu	laufen
zurück	**legen**
vor	**legen**
hin	**legen**
ab	**legen**
auf	**legen**
ab	lösen
aus	lösen
auf	lösen
vor	**nehmen**
hin	**nehmen**
ein	**nehmen**
ab	**nehmen**
zu	**nehmen**
auf	**nehmen**
an	**nehmen**
hinüber	nicken
um	**reißen**
ab	**reißen**
auf	**reißen**
zurück	rennen
hin	rennen
zu	rennen

zurück	**rufen**
hinüber	**rufen**
zu	**rufen**
aus	**rufen**
an	**rufen**
vor	sehen
um	sehen
hin	sehen
ein	sehen
hinüber	sehen
zu	sehen
aus	sehen
auf	sehen
an	sehen
hin	**setzen**
ein	**setzen**
ab	**setzen**
auf	**setzen**
an	**setzen**
ein	schalten
ab	schalten
aus	schalten
hinüber	**schauen**
um	**schauen**
hin	**schauen**
zu	**schauen**
an	**schauen**
hinüber	schieben
zurück	schieben
vor	schieben
hin	schieben
zu	schieben
hinüber	**schielen**
vor	schlagen
um	schlagen
ein	schlagen
zu	schlagen
auf	schlagen
ein	**schließen**
ab	**schließen**
auf	**schließen**
an	**schließen**

Fortsetzung auf nächster Seite

zurück	schreien	vor	strecken	**hinüber**	**winken**
zu	schreien	hin	strecken	**zurück**	**winken**
auf	schreien	aus	strecken	**ab**	**winken**
an	schreien	**hinüber**	**tragen**	**zu**	**winken**
hinüber	**springen**	zurück	treten	hinüber	zeigen
zurück	**springen**	vor	treten	**zurück**	**ziehen**
ab	**springen**	ein	treten	**vor**	**ziehen**
auf	**springen**	**um**	**wenden**	**um**	**ziehen**
hinüber	starren	**ein**	**wenden**	**hin**	**ziehen**
hinüber	**stehen**	**ab**	**wenden**	**ein**	**ziehen**
um	steigen	**zu**	**wenden**	**aus**	**ziehen**
ein	steigen	zurück	werfen	**hinüber**	**ziehen**
ab	steigen	vor	werfen	**an**	**ziehen**
aus	steigen	um	werfen		
auf	steigen	hin	werfen		
zurück	**stoßen**	ein	werfen		
um	**stoßen**	hinüber	werfen		
ab	**stoßen**	zu	werfen		
aus	**stoßen**				
auf	**stoßen**				
an	**stoßen**				

Adverbien mit der höchsten Frequenz (alphabetisch):

1. allein
2. allerdings
3. also
4. andererseits
5. anders
6. anfangs
7. auch
8. außerdem
9. bald
10. beispielsweise
11. bekanntlich
12. bereits
13. besonders
14. bestens
15. bestimmt
16. bisher
17. bislang
18. bisschen
19. da
20. damals
21. dann
22. denn
23. dennoch
24. diesmal
25. doch
26. dort
27. eben
28. ebenfalls
29. ebenso
30. ehemals
31. eher
32. eigentlich
33. einfach
34. endlich
35. erst
36. erstens
37. etwas
38. fast
39. freilich
40. früher
41. ganz
42. gar
43. genauso
44. gerade
45. gern
46. gestern
47. halbwegs
48. her
49. heute
50. hier
51. hin
52. hinten
53. immer
54. insgesamt
55. inzwischen
56. jedoch
57. jetzt
58. jeweils
59. kaum
60. lange
61. mal
62. manchmal
63. mehr
64. meist
65. mindestens
66. nämlich
67. natürlich
68. nie
69. noch
70. nun
71. nur
72. oft
73. rechts
74. rund(um)
75. sehr
76. selbst
77. schließlich
78. schon
79. so
80. sofort
81. sogar
82. sonst
83. soweit
84. tatsächlich
85. überhaupt
86. unten
87. viel
88. vielleicht
89. weiterhin
90. wenig
91. wenigstens
92. wieder
93. wohl
94. zudem
95. zuletzt
96. zumindest
97. zunächst
98. zurück
99. zusammen
100. zwar

Obwohl diese Liste ein Ergebnis des Durchschnittsvorkommens der häufigsten Adverbien in gleich drei Korpora (DeReWo = DeReKo, InterCorp (Kernkorpus) und deTenTen) darstellt, ist ein Zweifel an der Richtigkeit der Berechnung angebracht: viele Adverbien sind homonym zu Präpositionen, Konjunktionen oder Partikel (z.B. *denn*).

Präpositionen (nach Frequenz):

1. in
2. mit
3. auf
4. von
5. an
6. zu
7. für
8. aus
9. nach
10. über
11. vor
12. bei
13. durch
14. um
15. unter
16. gegen
17. ohne
18. zwischen
19. hinter
20. seit

Die häufigsten Präpositionen (alphabetisch):

ab
an
angesichts
auf
aufgrund
aus
außer
außerhalb
bei
bis
dank
durch
einschließlich
entlang
für
gegen
gegenüber
gleich
hinter
in
infolge
inmitten
innerhalb
je
jenseits
laut
mit
mitsamt
nahe
nach
namens
neben
oberhalb
ohne
per
pro
samt
seit
statt
trotz
über
um
unter
unterhalb
von
vor
während
wegen
zu
zugunsten
zwischen

Interjektionen (nach Frequenz):

Internettexte: Korpus deTenTen

1. ja
2. Na
3. naja
4. Ach/ach
5. hallo
6. Oh
7. Tja
8. nu
9. Ah
10. äh
11. He
12. Mann
13. Aha
14. ha
15. Bravo
16. Huch
17. stop
18. Basta
19. ade
20. zack
21. au
22. hurra
23. hopp
24. piep
25. schwupp
26. Gell
27. Ätsch
28. tschüß
29. Hoho
30. prost
31. schwupps
32. autsch
33. uh
34. marsch
35. ahoi

Belletristische Texte: Korpus InterCorp_de

1. Na	8. nu	15. ha
2. Ach	9. he	16. Bravo
3. Oh	10. marsch	17. ade
4. au	11. basta	18. hurra
5. Hallo	12. hopp	19. Uh
6. Ah	13. Menschenskind	20. Prost
7. Aha	14. stop	21. Bah

Die häufigsten Interjektionen (alphabetisch):

Geschriebenes Deutschen (allgemein)

ade	Gell	Na
Ah	ha	naja
äh	Hallo	nu
Aha	he	Oh
ahoi	Hoho	piep
Ach	hopp	Prost
Ätsch	Huch	schwupp(s)
au	hurra	stop
autsch	ja	Tja
Bah	Mann	tschüß
basta	marsch	Uh
Bravo	Menschenskind	zack

Bemerkung:
Prinzipiell werden Interjektionen kleingeschrieben. Die Groß-/Kleinschreibung in den vorliegenden Listen bezeichnet die häufigere Form in Korpora und deutet darauf hin, ob die Präposition einen selbstständigen Satz bildet (bzw. am Anfang eines Satzes steht) oder ob sie in einem Satz integriert ist.

7.3 Satz- und Textdiakritika (nach Häufigkeit):

1. ,	7. ;	13. …
2. .	8. » «	14. /
3. ( )	9. ?	15. =
4. "	10. !	16. --
5. -	11. [ ]	17. `
6. :	12. '	18. { }

8. Korpusabfragehilfen

8.1 Abfragbare Affixe

Abfrage im DeReKo

Im **DeReKo** können Affixe folgendermaßen ausgesucht werden:

- **Einstellung der Suchoptionen:** Lemmatisierung → sonstige Wortformen auswählen
- **Suchfeldeingabe Präfixe:** &ab- →
- **Suchfeldeingabe Suffixe:** &-abel →
- Ergebnisse → Export
- **Export der Wortformen:** Expansionsliste auswählen

Die Liste suchbarer Affixe in COSMAS $II_{win,\ web}$ (2012) ist über das Hilfeportal[56] abrufbar.

Präfixe:	*ab-, abwärts-, agrar-, agro-, all-, alt-, an-, anti-, aqua-, äqui-, ar-, archä-, auf-, aus-, be-, bei-, bi-, co-, d-*[1]*, da-, dar-, de-, dein-, di-, ein-, ent-, epi-, ex-, gar-, ge-, gegen-, geo-, gut-, häm-, heim-, hell-, hemi-, her-, hin-, irr-, jen-, ko-, los-, mit-, neo-, non-, ober-, para-, per-, phen-, post-, prä-, pre-, re-, rück-, rum-, semi-, sub-, supra-, teil-, über-, um-, un-, unter-, ur-, ver-, vor-, weg-, wider-, wieder-, zer-, zu-* (insgesamt 72)
Suffixe:	*-abel, -achsig, -al, -artig, -bar, -er, -fach, -förmig, -getreu, -haft, -haltig, -heit, -ig, -in, -isch, -isierung, -ismus, -ist, -istin, -ität, -itis, -jährig, -kalibrig, -keit, -lagig, -ler, -lich, -los, -mal, -maschig, -mässig*[2]*/-mäßig, -phasig, -reihig, -sam, -schaft, -schaftlich, -seitig, -strängig, -teilig, -tum, -ung, -volumig, -weise, -wertig, -zylindrig* (insgesamt 45) *-chen, -lein* (Diminutivsuffixe)

Anmerkungen
[1] Artikel und Artikelwörter beginnend auf *d-*
[2] Für Recherchen v.a. in schweizerischen Texten

Abfrage im InterCorp

Im **InterCorp_de** können Affixe nur einzeln ausgesucht werden.

Beispiel: Präfix *ab-*:

- **Query Type:** CQL → **CQL:** [lemma="ab.*"] → Frequency → Node forms

Dann müssen einzelne Formen aussortiert werden.

[56] http://www.ids-mannheim.de/cosmas2/web-app/hilfe/suchanfrage/affixe.html

8.2 Tag-Kürzel

Die folgenden Kürzel der morphosyntaktischen Annotation werden in den meisten Korpora mit Deutsch verwendet, also auch im **DWDS**, **DeReKo** (Archiv TAGGED-T) und im **InterCorp_de**.

Das Original der folgenden Tabelle ist als **STTS Tag Table** (1995/1999) (Schmid 1995) auf der Homepage des Instituts für Maschinelle Sprachverarbeitung, Universität Stuttgart zugänglich (siehe IMS (2013) im Literaturverzeichnis). Hier ist sie leicht modifiziert und ergänzt. Beispiele der KWICs sind in Fettschrift gesetzt.

Alphabetische Auflistung des Tag-Sets (STTS)

Tag	Beschreibung	Beispiel
ADJA	attributives Adjektiv	das **große** Haus
ADJD	adverbiales oder prädikatives Adjektiv	er fährt **schnell**, er ist **schnell**
ADV	Adverb	**schon**, **bald**, **doch**
APPO	Postposition	ihm **zufolge**, der Sache **wegen**
APPR	Präposition; Zirkumposition links	**in** der Stadt, **ohne** mich
APPRART	Präposition mit Artikel	**im** Haus, **zur** Sache
APZR	Zirkumposition rechts	von jetzt **an**
ART	bestimmter oder unbestimmter Artikel	**der**, **die**, **das**, **ein**, **eine**
CARD	Kardinalzahl	**zwei** Männer, im Jahre **1994**
FM	fremdsprachliches Material	Er hat das mit "**A big fish**" übersetzt
ITJ	Interjektion	**mhm**, **ach**, **tja**
KOKOM	Vergleichskonjunktion	**als**, **wie**
KON	nebenordnende Konjunktion	**und**, **oder**, **aber**
KOUI	unterordnende Konjunktion mit "zu" und Infinitiv	**um** zu leben, **anstatt** zu fragen
KOUS	unterordnende Konjunktion mit Satz	**weil**, **dass**, **damit**, **wenn**, **ob**
NE	Eigennamen	**Hans**, **Hamburg**, **HSV**
NN	normales Nomen	**Tisch**, **Herr**, das **Reisen**
PAV	Pronominaladverb	**dafür**, **dabei**, **deswegen**, **trotzdem**
PDAT	attribuierendes Demonstrativpronomen	**jener** Mensch
PDS	substituierendes Demonstrativpronomen	**dieser**, **jener**
PIAT	attribuierendes Indefinitpronomen ohne Determiner	**kein** Mensch, **irgendein** Glas
PIDAT[1]	attribuierendes Indefinitpronomen mit Determiner	~~ein **wenig** Wasser, die **beiden** Brüder~~[1]
PIS	substituierendes Indefinitpronomen	**keiner**, **viele**, **man**, **niemand**
PPER	irreflexives Personalpronomen	**ich**, **er**, **ihm**, **mich**, **dir**
PPOSAT	attribuierendes Possessivpronomen	**mein** Buch, **deine** Mutter
PPOSS	substituierendes Possessivpronomen	**meins**, **meinige**, **deiner**, **ihrs**
PRELAT	attribuierendes Relativpronomen	der Mann, **dessen** Hund

Fortsetzung auf nächster Seite

PRELS	substituierendes Relativpronomen	der Hund, **der**
PRF	reflexives Personalpronomen	**sich**, **einander**, **dich**, **mir**
PTKA	Partikel bei Adjektiv oder Adverb	**am** schönsten, **zu** schnell
PTKANT	Antwortpartikel	**ja**, **nein**, **danke**, **bitte**
PTKNEG	Negationspartikel	**nicht**
PTKVZ	abgetrennter Verbzusatz	er kommt **an**, er fährt **rad**
PTKZU	"zu" vor Infinitiv	**zu** gehen
PWAT	attribuierendes Interrogativpronomen	**welche** Farbe, **wessen** Hut
PWAV	adverbiales Interrogativ- oder Relativpronomen	**warum**, **wo**, **wann**, **worüber**, **wobei**
PWS	substituierendes Interrogativpronomen	**wer**, **was**
TRUNC	Kompositions-Erstglied	**Sumpf-** und Seelandschaft
VAFIN	finites Hilfsverb (auxiliar)	du **bist**, wir **werden**
VAIMP	Imperativ eines Hilfsverbs[2]	**sei** ruhig!
VAINF	Infinitiv eines Hilfsverbs	**werden**, **sein**, **haben**
VAPP	Partizip Perfekt eines Hilfsverbs	**gewesen**, **worden**
VMFIN	finites Modalverb	**kann**, **sollte**, **mag**
VMINF	Infinitiv eines Modalverbs	**können**, **müssen**, **wollen**, **sollen**, **dürfen**, **mögen**[3]
VMPP	Partizip Perfekt eines Modalverbs	**gekonnt**, er hat gehen **können**
VVFIN	finites Vollverb	du **gehst**, wir **kommen** an
VVIMP	Imperativ eines Vollverbs	**komm** !
VVINF	Infinitiv eines Vollverbs	**gehen**, **ankommen**
VVIZU	Infinitiv eines Vollverbs mit "zu''	**anzukommen**, **loszulassen**
VVPP	Partizip Perfekt eines Vollverbs	**gegangen**, **angekommen**
XY	Nichtwort, Sonderzeichen enthaltend	**3:7**, **H2O**, **D2XW3**
$(	sonstige Satzzeichen	**- ,()**
$,	Beistrich/ Komma	**,**
$.	satzbeendende Interpunktion	**. ? ! ; :**

Anmerkungen

[1] Kürzel nur im DWDS vorhanden und liefert Ergebnisse: *manch* (*ein*), *welch* (*ein*), *solch* (*ein*), *all* (*die*).

[2] Imperative von *sein* und *werden* in der Rolle des Kopulaverbs. Sinnvolle Ergebnisse liefert nur das DWDS.

[3] Frequenz in dieser Reihenfolge

Thematische Auflistung der morphosyntaktischen Tags

Die folgende Tabelle bringt die Übersicht über einzelne Kategorien, die in drei morphosyntaktisch annotierten Korpora abgefragt werden können: **DeReKo** (TAGGED-M, TAGGED-T) und **InterCorp**_de. (In normaler Schrift angeführte Tags in der Spalte von InterCorp gelten natürlich auch für das DWDS.)

Abfrage im DeReKo

Die Abfrageerstellung erfolgt über die sog. „graphische Suche“: aus dem Menüangebot werden die Kategorien hierarchisch ausgewählt. Übergeordnete Ebenen („Oberklassen“) sind in Fettschrift gesetzt.

Abfrage im InterCorp

Die Abfrageerstellung erfolgt über

- **Query Type:** CQL → **CQL:** [tag="…"].

Die Abfrage muss gewöhnlich durch die Platzhalter (.*) ergänzt werden, um alle zutreffenden Belege abzurufen (z.B. [tag="ADJ.*"]).

DeReKo			**InterCorp**
Archiv: TAGGED-M (MECOLB)		**Archiv: TAGGED-T (TreeTagger)**	**InterCorp_de TreeTagger/ *RFTagger*[57]**
Substantiv	Unterklasse: Gattungsbez. Eigenname Kasus Numerus Genus	**Nomina:** normale Eigennamen	N NN *N.Reg.* NE *N.Name* *Nom., Gen., Dat., Acc.* *Sg., Pl.* *Masc, Fem, Neut*
Adjektiv	 Kasus Numerus Genus Komparation Form	**Adj.:** attributiv prädikativ oder adverbial	ADJ ADJA ADJD *Nom., Gen., Dat., Acc.* *Sg., Pl.* *Masc, Fem, Neut* *Pos., Comp., Sup.* -
Pronomen	Art: demonstrativ - interrogativ personal - reflexiv relativ Kasus: Numerus: Genus:	**Pron.:** Demonstrativpronomen Indefinitpronomen Interrogativpronomen Personalpronomen Possessivpronomen Reflexivpronomen Relativpronomen	*PRO* *Dem.* *Indef.* *Inter.* *Pers.* *Poss.* *Refl.* *Rel.* *Nom., Gen., Dat., Acc.* *Sg., Pl.* *Masc, Fem, Neut*

Fortsetzung auf nächster Seite

[57] Schmid/ Laws 2008. Tags in *Kursivschrift* ab der Version 7.

DeReKo			**InterCorp**
Artikel	Form bestimmt unbestimmt Kasus Numerus Genus	**Art.**	ART *Def.* *Indef.* *Nom., Gen., Dat., Acc.* *Sg., Pl.* *Masc, Fem, Neut*
Numerale	Typ Kardinale Ordinale Bruchzahl Kasus Numerus Genus	**Zahlen**	 CARD - - - - -
Verb		**Verb** finit ohne Imperativ auxiliar voll modal Infinitiv auxiliar voll modal mit zu Partizip Perfekt auxiliar voll modal	V VFIN VAFIN *VFIN.Aux* VVFIN *VFIN.Full* VMFIN *VFIN.Mod* VINF VAINF *VINF.Aux.* VVINF *VINF.Full* VMINF *VINF.Mod.* VVIZU *VINF.Full.zu* VPP VAPP *VPP.Aux* VVPP *VPP.Full* VMPP *VPP.Mod.*
	Genus Verbi Tempus (Präsens) (Präteritum) Modus (Indikativ) (Konjunktiv) (Imperativ) Numerus (Singular) (Plural) Person (1.) (2.) (3.)		- V *Pres.* *Past.* V *Ind.* *Subj.* VIMP V *Sg.* *Pl.* V *1.* *2.* *3.*
Verbzusatz		(siehe Partikel)	PTKVZ
Adverb		**Adv.**	ADV
Konjunktion	Typ subordinierend koordinierend	**Konj.** unterordnende mit Infinitiv mit Satz nebenordnende Vergleichspartikel (*als, wie*)	K KOU KOUI KOUS KON KOKOM

Fortsetzung auf nächster Seite

Präposition	Position Präposition Postposition Kasus Artikelverschmelzung	**Adp.** Präpositionen Postposition Zirkumposition rechts (*auf … **hin***) Präpositionen mit Artikel	AP APPR APPO *Gen., Dat., Acc*. APZR APPRART
(Partikel)		**Partikel** *zu* vor Infinitiv Negationspartikel abgetrennter Verbzusatz (*all*)*zu* bei Adj. oder Adverb Antwortpartikel	PTK PTKZU PTKNEG PTKVZ PTKA PTKANT *PART.Ans*
(Sonstiges)		**Sonst.** Interjektionen Kompositions-Erstglied Nichtwörter fremdsprachliches Material Satzzeichen, Kürzel	 ITJ TRUNC *Adj., Noun., Verb* XY FM *SYM*

8.3 Tastenkürzel, Shortcuts

Die Tastenkürzel funktionieren zuverlässig im MS-Office-Paket auf den deutschen Standardtastaturen (de) für Deutsch (Deutschland), (Österreich), (Schweiz), (Liechtenstein) und (Luxemburg) bzw. auf der tschechischen und größtenteils auch auf der slowakischen Standardtastatur.
Die Tastenkombinationen (shortcuts) gelten allgemein.

Zeichen	Beschreibung	Tastenkürzel		Tastenkombination (shortcut)
		Tastatur		
		deutsche (de)	tschechische (cs)	
(	runde Klammer auf	shift+8	shift+)	Alt+40
)	runde Klammer zu	shift+9	)	Alt+41
{	geschwungene/ geschweifte Klammer auf	AltGr+7	AltGr+b	Alt+123
}	geschwungene/ geschweifte Klammer zu	AltGr+0	AltGr+n	Alt+125
|	senkrechter Abgrenzungsbalken	AltGr+>	AltGr+w	Alt+0124
[	eckige Klammer auf	AltGr+8	AltGr+f	Alt+91
]	eckige Klammer zu	AltGr+9	AltGr+g	Alt+93
/	Schrägstrich/ Querstrich/ slash	shift+7	shift+ů	Alt+47
****	back slash/ umgekehrter Schrägstrich	AltGr+ß	AltGr+q	Alt+92
"	Anführungszeichen ("Gänsefüßchen")	shift+2	shift+ů	Alt+34
«	Anführungszeichen "Guillemets"			Alt+0171
»	Anführungszeichen "Guillemets"			Alt+0187
‹	Anführungszeichen "Guillemets" (einfach)			Alt+0139
›	Anführungszeichen "Guillemets" (einfach)			Alt+0155
^	Zirkumflex	^	AltGr+3+space	Alt+94
<	größer als	<	AltGr+.	Alt+60
>	kleiner als	>	AltGr+,	Alt+62
*****	Asterisk/ Sternchen	shift++	AltGr+-	Alt+42
+	Pluszeichen	+	+	Alt+43
-	Minuszeichen	-	-	Alt+45
#	Raute	#	AltGr+x	Alt+35
%	Prozent	shift+4	shift+=	Alt+37
@	At-Zeichen	AltGr+q	AltGr+v	Alt+64
!	Rufzeichen	shift+1	shift+§	Alt+33
?	Fragezeichen	shift+ß	shift+,	Alt+63
&	Ampersand	shift+6	AltGr+c	Alt+38
$	Dollarzeichen	shift+4	AltGr+ů	Alt+36
ß	scharfes S	ß	AltGr+§	Alt+225
ä	a-Umlaut	ä	¨+a	Alt+132
Ä	A-Umlaut	Ä	¨+shift+a	Alt+142
ö	o-Umlaut	ö	¨+o	Alt+148
Ö	O-Umlaut	Ö	¨+shift+o	Alt+153
ü	u-Umlaut	ü	¨+u	Alt+129
Ü	U-Umlaut	Ü	¨+shift+u	Alt+154

8.4 InterCorp: CQL-Abfragen

In den folgenden Tabellen sind Syntaxhilfen für einige Abfragen im InterCorp angeführt. Alle Abfragen gelten nur für den Abfragemodus CQL (**Query Type: CQL**). Die Reihung in der ersten Tabelle (Reihung nach Abfrage: S. 211-214) geht von den einfachsten Abfragen nach Wort und Wortteilen zu den kompliziertesten Abfragen nach Kombinationen von Elementen im Satz.

Die zweite Tabelle (Alphabetische Reihung: S. 215-218) entspricht der ersten, die Reihung der Abfragen ist jedoch nach der Bedeutung der Abfrage angeordnet und geht alphabetisch von *Abstand* bis zu *Wortform*. Die offenbare Mischung von grammatikalischen Kategorien (z.B. *Partizipialphrase*) und korpuslinguistischen Abfragebegriffen (z.B. *Abstand*) darf nicht verwirren. Sie ist durch die Abfragemöglichkeiten des Korpusmanagers verursacht. Die grammatikalischen und sonstigen linguistischen Kategorien können nämlich größtenteils nicht direkt abgefragt werden, sondern sie müssen über gut überlegte Kombinationen von möglichen Suchfeldeingaben abgefragt werden (z.B. die Abfrage nach einer Partizipphrase mit Partizip I kann über eine Kombination von einem Tag für Adjektive in der Attributposition (ADJA) und einem Lemma, das auf *-end* endet, erfolgen).

Eine vollständige Tabelle für die Abfragen kann es wegen den unzähligen Kombinationsmöglichkeiten der abzufragenden Elemente nicht geben. Die folgenden Tabellen sind ein Bruchteil aller Abfragemöglichkeiten. Sie sind als erste Hilfe für die üblichsten Abfragen konzipiert.

CQL-Abfragen (InterCorp): Reihung nach Abfrage

Abfrage	**Bedeutung**	**Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)**
[word="rund"]	Wortform	**rund**
[word="rund.*"]	Wortanfang	**rund, runde, rundherum, Runde, rundlich**
[word=".*rund"]	Wortende	**aufgrund, Grund, rund, Hintergrund**
[word="Schlu(ß|ss)"]	Alternative *ß* oder *ss* im Wort	**Schluss, Schluß**
[word="gr[^ü]ße"]	Alternative (*ü* ignorieren)	**große, Größe**
[wort="m[^a|ü]de"]	Alternative (*a* oder *ü* ignorieren)	**Mode**
[lemma="Grund"]	Grundform/ Lemma	**Gründe, Grund, Grunde, Grundes**
[lemma=".*schaft"&!lemma="Gemeinschaft"]	Wort auf *-schaft*, jedoch NICHT Gemeinschaft	**Wirtschaft, Gesellschaft, Landwirtschaft**
[tag="VVINF.*"] [word="hätte"] [tag="VMINF.*"]	Verbalkomplex („österreichische“ Reihenfolge)	wie ich sie **auffassen hätte können**; Mehrwert , den er **erbringen hätte können**; etwas stärker **hervorheben hätte müssen**; einer wirklichen Selbstkritik **unterziehen hätte sollen**
[word="hätte"] [tag="VVINF.*"] [tag="VMINF.*"]	Verbalkomplex („deutsche“ Reihenfolge)	… , daß sie seine Mutter **hätte sein können**.; , der genau **hätte sagen können**,; … die Preise , die ich ihnen **hätte zahlen müssen**
[tag="N.*"&lemma=".*el"]	Substantiv auf-*el*	**Artikel, Mittel, Ziel, Beispiel…**
[tag="N.*"&lemma=".*el"&!lemma="Artikel"]	Substantiv auf-*el*, außer *Artikel* (Fast 50% aller Treffer der Subst. auf *-el* fallen auf das Wort *Artikel*.)	**Mittel, Ziel, Beispiel, Handel, Kapitel** … aber auch Desubstantive auf *-el*: **Geflügel, Flügel, Gürtel, Rätsel, Bündel** …
[tag="APPR"&word="auf"]	Präposition *auf*	**auf dem Bauch; auf dem Boden; auf den Boden**
[tag="PTKVZ"&word="auf"] [word="\?"]	Verbzusatz *auf* am Ende eines Fragesatzes	Nehmen Sie das **auf ?**; Warum regst du dich denn so **auf ?**
[lemma="fordern"][]*[word="auf"] within <s />	Verb mit abgetrenntem Zusatz	**fordert** die Kommission/ den Rat (nachdrücklich) **auf**; **fordere** Sie **auf**, **forderte** ihn **auf**, **fordert** dazu **auf**

Abfrage	**Bedeutung**	**Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)**
[lemma="auffordern"]	Verb mit Zusatz	**auffordern, aufzufordern, auffordert, aufforderte, aufforderten, auffordere**
[word="frage"] []* [word="ich"] within <s/>	Satzelemente: inverse Wortstellung (*frage … ich*)	**frage ich; frage mich, ob ich; frage ich mich, ob ich**
[word="ich"][][word="frage"]	Satzelemente: normale Wortstellung (*ich … frage*; dazwischen 1 Wort)	**ich mich/ Sie/ ihn frage; ich freundlich frage**
[lemma=".*park"]within <div·author="Bachmann,·Ingeborg"/>	Grundformen/ Lemma aller Wörter auf *-park* in Werken von *Bachmann, Ingeborg*	**Stadtpark, Stadtparks, Resselpark**
[lemma="lieb.*"]within <div·title="Scherz"/>	Grundformen/ Lemmata beginnend mit *lieb-* im Werk *Scherz* (von Milan Kundera)	**liebte, geliebt, lieben, lieber, liebste, Liebe**
[lemma="[Ww]ien.*"]within	Grundformen/ Lemmata beginnend mit *Wien-* oder *wien-* im Werk *Malina* (von Ingeborg Bachmann)	**Wien; Wiener; wienerisches; Wienerwald; Wienerinnen**
[word="allein"][word="\."]	Satzende: Wort *allein* am Satzende	**allein.**
[word=".*lein"&!word="[k|K]lein|[a|A]llein"]	Negation: alle Wörter auf *-lein*, jedoch nicht *Klein*/*klein* und *Allein*/*allein*	**Fräulein, Häuflein, Büchlein, Männlein, Henlein, Äuglein, Bäuchlein, Bächlein…**
<s>[word="Dich"]	Satzanfang: *Dich* am Satzanfang	**Dich** aber beneide ich; **Dich** kann nichts rühren; **Dich** so zu erschrecken; **Dich** zu lieben bedeutet
<s>[tag="VMFIN"]	Satzanfang: Modalverb, finite Form am Satzanfang	**Kann** ich Ihnen helfen?; **Sollte** er das erzählen?; **Können** Sie bitte erklären,
<s>[tag="VMFIN"&lemma="mögen"]	Satzanfang: Modalverb *mögen* am Satzanfang	**Mag, Möge, Mögen, Möchten…**
[lemma="Kopf"][]{0,5}[lemma="Hand"]	Abstand: Lemma *Kopf* und Lemma *Hand* im Abstand von maximal 5 Positionen	**Kopf in die Hände; Kopf in beide Hände; Kopf mit beiden Händen; Kopf und die Hände**
[lemma="Ku(ss|ß)"][]*[lemma="Liebe"] within <p/>	Abstand: Lemma *Kuss* oder *Kuß* und Lemma *Liebe* in einem Absatz	**stoßweise gehender Atem und seine wilden Küsse? Was war das alles, wenn es keine Liebe war?**

Abfrage	**Bedeutung**	**Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)**
[lemma="essen"][]*[word="gern"] within <s/>	Abstand: Lemma *essen* und Wort *gern* in einem Satz	**Essen Sie gern Gorgonzola?**
[lemma="halten"] []* [word="fest"] within <s id=".*"/>	Verb mit abgetrenntem Zusatz	ich **hielt** sie an den Schultern **fest;** Er blieb stehen und **hielt** sich am Geländer **fest**.
[lemma="halten"] []* [word="fest"] [tag="\$."]	Abstand: Lemma *halten* im beliebigem Abstand vom Wort *fest* am Satzende	**hält** die Kommission **fest**.
[lemma="halten"] []* [word="fest"] [tag="\$."] within <s id=".*"/>	Verb mit abgetrenntem Zusatz (in einem Satz); Zusatz am Satzende	aber kaum **hielt** er sich derart **fest**,; Für einen kurzen Augenblick **hielt** ihr Blick den seinen **fest**.
[word="es"] []*[lemma="verlangen"] []{0,3} [word="nach"] within <s/>	Phrase: jmdn. *verlangt es nach* etw. (Reihenfolge im Satz: *es - verlangen - nach*)	**es** hat mich so **verlangt nach**; **es** satt , und ihr Geist **verlangte nach**
[lemma="verlangen"][]*[word="es"] []{0,3} [word="nach"] within <s/>	Phrase: jmdn. *verlangt es nach* etw. (Reihenfolge im Satz: *verlangen - es - nach*)	Die Männer **verlangte es nach** weiblicher Gesellschaft; **verlangt es den Menschen mitunter nach** dem monotonen Rhythmus
<s/> containing [lemma="fest"] containing [lemma="halten"]	Satz mit Lemmata *fest* und *halten* (Als KWIC erscheint der ganze Satz.)	**Sie hielten mich erst in Bagram fest, dann in Kandahar und schließlich in Guantánamo Bay. Er hob die Hände, ertastete sich eine Stuhllehne und hielt sich fest, und wir sprachen kein Wort mehr miteinander, bis Petra angezogen herunterkam, mit Mantel und Kopftuch.**
[tag="\$."]	Satzende (beliebiges Zeichen am Satzende)	**. ,] " :**
[tag="\$."&word="\."]	Satzende: Punkt am Ende	**.**
[word="\."]	Satzzeichen: Punkt	**.**
[word="\»"]	Satzzeichen: Anführungszeichen "Guillemets"	**»**

Abfrage	Bedeutung	Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)
[word="\""]	Satzzeichen: Anführungszeichen "Gänsefüßchen"	**"**
[word="\?"]	Satzzeichen: Fragezeichen	**?**
[word="\!"]	Satzzeichen: Rufzeichen	**!**
[word="\!"] [word="\?"]	Satzzeichen: Rufzeichen und Fragezeichen	"Dann gibst du mir also recht**!?**"
<s/> containing [tag="VVFIN"] containing [tag="PTKVZ"]	Satz mit einem Vollverb in finiter Form, in demselben Satz auch ein Verbzusatz (Als KWIC erscheint der ganze Satz.)	**Der Vorsitzende nimmt an der Abstimmung nicht teil. Es ist eines der drei Stücke, die Mama auf dem Klavier spielen kann, geschrieben von dem Komponisten Polívka, ich höre zu, und mein Zorn vergeht.**
[tag="APPR"] [tag="ADJA"&lemma=".*end"]	Partizipphrase (mit Partizip I) nach einem Artikel	Sie holt die Chefin, Charlotte, die stürzt **in fliegender** Eile hinauf.
[tag="APPR"] [tag="ADJA"&lemma=".*(en|t)"]	Partizipphrase (mit Partizip II) nach einem Artikel	Erzeugnisse, die **in unverändertem** Zustand ausgeführt werden
[tag="ART"] [tag="ADJA"] [tag="NE"]	Eigenname mit Artikel und Attribut	**die falsche Kleopatra** thronte immer noch auf ihrem dunkelroten Kißchen; Aber **die blonde Käthe** stolperte; Voraussetzung für die Schaffung **eines integrierten Europa**
[word="des"] [tag="NE"]	Eigenname nach *des* (Maskulin im Genitiv)	der Sprachen **des Balkans**; die Anerkennung **des Kosovo**; Die Wirklichkeit überholte auch die tschechische Utopie **des Karel** Čapek; unter der Statue **des Jan** Nepomuk,
[word="des"] [tag="ADJA"] [tag="NE"&!word=".*s"]	Eigenname (im Genitiv) ohne *-s* mit bestimmten Artikel *des* und Attribut	Feudalist und Verehrer **des alten Österreich**; Blicke waren nach oben auf das Reiterstandbild **des heiligen Wenzel** gerichtet

CQL-Abfragen (InterCorp): Alphabetische Reihung nach Bedeutung der Abfrage

Bedeutung	Abfrage	Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)
Abstand: Lemma *essen* und Wort *gern* in einem Satz	[lemma="essen"][]*[word="gern"] within <s/>	**Essen Sie gern Gorgonzola?**
Abstand: Lemma *halten* im beliebigem Abstand vom Wort *fest* am Satzende	[lemma="halten"] []* [word="fest"] [tag="\$."]	**hält** die Kommission **fest.**
Abstand: Lemma *Kopf* und Lemma *Hand* im Abstand von maximal 5 Positionen	[lemma="Kopf"][]{0,5}[lemma="Hand"]	**Kopf in die Hände; Kopf in beide Hände; Kopf mit beiden Händen; Kopf und die Hände**
Abstand: Lemma *Kuss* oder *Kuß* und Lemma *Liebe* in einem Absatz	[lemma="Ku(ss|ß)"][]*[lemma="Liebe"] within <p/>	**stoßweise gehender Atem und seine wilden Küsse ? Was war das alles, wenn es keine Liebe war?**
Alternative: *Schluß* oder *Schluss*	[word="Schlu(ß|ss)"]	**Schluss, Schluß**
Alternative (*a* oder *ü* ignorieren)	[wort="m[^a|ü]de"]	**Mode**
Alternative (*ü* ignorieren)	[word="gr[^ü]ße"]	**große, Größe**
Eigenname (im Genitiv) ohne *-s* mit bestimmten Artikel *des* und Attribut	[word="des"] [tag="ADJA"] [tag="NE"&!word=".*s"]	Feudalist und Verehrer **des alten Österreich**; Blicke waren nach oben auf das Reiterstandbild **des heiligen Wenzel** gerichtet
Eigenname mit Artikel und Attribut	[tag="ART"] [tag="ADJA"] [tag="NE"]	**die falsche Kleopatra** thronte immer noch auf ihrem dunkelroten Kißchen; Aber **die blonde Käthe** stolperte; Voraussetzung für die Schaffung **eines integrierten Europa**
Eigenname nach *des* (Maskulina im Genitiv)	[word="des"] [tag="NE"]	der Sprachen **des Balkans**; die Anerkennung **des Kosovo**; Die Wirklichkeit überholte auch die tschechische Utopie **des Karel** Čapek ; unter der Statue **des Jan** Nepomuk ,
Grundform/ Lemma	[lemma="Grund"]	**Gründe, Grund, Grunde, Grundes**

Bedeutung	Abfrage	Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)
Grundform/ Lemma aller Wörter auf *-park* **in Werken** von *Bachmann, Ingeborg*	[lemma=".*park"]within <div·author="Bachmann,·Ingeborg"/>	**Stadtpark, Stadtparks, Resselpark**
Grundform/ Lemma beginnend mit *lieb-* **im Werk** *Scherz* (von Milan Kundera)	[lemma="lieb.*"]within <div·title="Scherz"/>	**liebte, geliebt, lieben, lieber, liebste, Liebe**
Grundform/ Lemma beginnend mit *Wien-* oder *wien-* **im Werk** *Malina* (von Ingeborg Bachmann)	[lemma="[Ww]ien.*"]within	**Wien; Wiener; wienerisches; Wienerwald; Wienerinnen**
Negation: alle Wörter auf *-lein*, jedoch nicht *Klein*/*klein* und *Allein*/*allein*	[word=".*lein"&!word="[k|K]lein|[a|A]llein"]	**Fräulein, Häuflein, Büchlein, Männlein, Henlein, Äuglein, Bäuchlein, Bächlein…**
Partizipphrase (mit Partizip I) nach einem Artikel	[tag="APPR"] [tag="ADJA"&lemma=".*end"]	Sie holt die Chefin, Charlotte, die stürzt **in fliegender** Eile hinauf.
Partizipphrase (mit Partizip II) nach einem Artikel	[tag="APPR"] [tag="ADJA"&lemma=".*(en|t)"]	Erzeugnisse , die **in unverändertem** Zustand ausgeführt werden
Phrase: jmdn. *verlangt es nach* etw. (Reihenfolge im Satz: *es - verlangen - nach*)	[word="es"] []*[lemma="verlangen"] []{0,3} [word="nach"] within <s/>	**es** hat mich so **verlangt nach**; **es** satt , und ihr Geist **verlangte nach**
Phrase: jmdn. *verlangt es nach* etw. (Reihenfolge im Satz: *verlangen - es - nach*)	[lemma="verlangen"][]*[word="es"] []{0,3} [word="nach"] within <s/>	Die Männer **verlangte es nach** weiblicher Gesellschaft; **verlangt es den Menschen mitunter nach** dem monotonen Rhythmus
Präposition *auf*	[tag="APPR"&word="auf"]	**auf dem Bauch; auf dem Boden; auf den Boden**
Satzanfang: *Dich* am Satzanfang	<s>[word="Dich"]	**Dich** aber beneide ich; **Dich** kann nichts rühren; **Dich** so zu erschrecken; **Dich** zu lieben bedeutet
Satzanfang: Modalverb *mögen* am Satzanfang	<s>[tag="VMFIN"&lemma="mögen"]	**Mag, Möge, Mögen, Möchten…**
Satzanfang: Modalverb, finite Form am Satzanfang	<s>[tag="VMFIN"]	**Kann** ich Ihnen helfen ?; **Sollte** er das erzählen ?; **Können** Sie bitte erklären ,

Bedeutung	Abfrage	Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)
Satz mit den **Lemmata** *fest* und *halten* (Als KWIC erscheint der ganze Satz.)	<s/> containing [lemma="fest"] containing [lemma="halten"]	**Sie hielten mich erst in Bagram fest, dann in Kandahar und schließlich in Guantánamo Bay. Er hob die Hände, ertastete sich eine Stuhllehne und hielt sich fest, und wir sprachen kein Wort mehr miteinander, bis Petra angezogen herunterkam, mit Mantel und Kopftuch.**
Satz mit einem Vollverb in finiter Form und mit einem Verbzusatz (Als KWIC erscheint der ganze Satz.)	<s/> containing [tag="VVFIN"] containing [tag="PTKVZ"]	**Der Vorsitzende nimmt an der Abstimmung nicht teil. Es ist eines der drei Stücke, die Mama auf dem Klavier spielen kann, geschrieben von dem Komponisten Polívka, ich höre zu, und mein Zorn vergeht.**
Satzelemente: inverse Wortstellung (*frage … ich*)	[word="frage"] []* [word="ich"] within <s/>	**frage ich; frage mich, ob ich; frage ich mich, ob ich**
Satzelemente: normale Wortstellung (*ich … frage*; dazwischen 1 Wort)	[word="ich"][][word="frage"]	**ich mich/ Sie/ ihn frage; ich freundlich frage**
Satzende (beliebiges Zeichen am Satzende)	[tag="\$."]	**. ,] " :**
Satzende: Punkt am Ende	[tag="\$."&word="\."]	**.**
Satzende: Wort *allein* am Satzende	[word="allein"][word="\."]	**allein .**
Satzzeichen: Anführungszeichen "Gänsefüßchen"	[word="\""]	**"**
Satzzeichen: Anführungszeichen "Guillemets"	[word="\»"]	**»**
Satzzeichen: Fragezeichen	[word="\?"]	**?**
Satzzeichen: Punkt	[word="\."]	**.**
Satzzeichen: Rufzeichen	[word="\!"]	**!**
Satzzeichen: Rufzeichen und Fragezeichen	[word="\!"] [word="\?"]	„ Dann gibst du mir also recht **! ?** “

Bedeutung	Abfrage	Ergebnis in Fettschrift, Reihenfolge nach Häufigkeit (wenn sinnvoll)
Substantiv auf *-el*	[tag="N.*"&lemma=".*el"]	**Artikel, Mittel, Ziel, Beispiel…**
Substantiv auf *-el*, außer *Artikel*	[tag="N.*"&lemma=".*el"&!lemma="Artikel"]	**Mittel, Ziel, Beispiel, Handel, Kapitel** … aber auch Desubstantive auf *-el*: **Geflügel, Flügel, Gürtel, Rätsel, Bündel** …
Verb mit abgetrenntem **Zusatz**	[lemma="fordern"][]*[word="auf"] within <s />	**fordert** die Kommission/ den Rat (nachdrücklich) **auf**; **fordere** Sie **auf**, **forderte** ihn **auf**, **fordert** dazu **auf**
Verb mit abgetrenntem **Zusatz** am Satzende	[lemma="halten"] []* [word="fest"] [tag="\$."] within <s ida=".*"/>	aber kaum **hielt** er sich derart **fest,** Für einen kurzen Augenblick **hielt** ihr Blick den seinen **fest.**
Verb mit Zusatz	[lemma="auffordern"]	**auffordern, aufzufordern, auffordert, aufforderte, aufforderten, auffordere**
Verbalkomplex („österreichische“ Reihenfolge)	[tag="VVINF.*"] [word="hätte"] [tag="VMINF.*"]	wie ich sie **auffassen hätte können**; Mehrwert , den er **erbringen hätte können**; etwas stärker **hervorheben hätte müssen**; einer wirklichen Selbstkritik **unterziehen hätte sollen**
Verbalkomplex („deutsche“ Reihenfolge)	[word="hätte"] [tag="VVINF.*"] [tag="VMINF.*"]	… , daß sie seine Mutter **hätte sein können**.; , der genau **hätte sagen können**,; … die Preise , die ich ihnen **hätte zahlen müssen**
Verbzusatz am Ende eines Fragesatzes	[tag="PTKVZ"&word="auf"] [word="\?"]	Nehmen Sie das **auf ?**; Warum regst du dich denn so **auf ?**
Wortanfang	[word="rund.*"]	**rund, runde, rundherum, Runde, rundlich**
Wiederholung von 3 bis 5 Adjektiven	[tag="ADJA "]{3,5}	um die ***neuen globalen politischen*** Realitäten anzuerkennen; In dem ***ausgezeichneten neuen britischen parlamentarischen*** Wahlkreis Daventry steht das Präzisionswerk…; die für die koordinierte Einführung eines ***europaweiten öffentlichen zellularen digitalen terrestrischen*** Mobilfunkdienstes in der Gemeinschaft bereitzustellen sind.
Wortende	[word=".*rund"]	**aufgrund, Grund, rund, Hintergrund**
Wort auf *-schaft*, jedoch NICHT Gemeinschaft	[lemma=".*schaft"&!lemma="Gemeinschaft"]	**Wirtschaft, Gesellschaft, Landwirtschaft**
Wortform	[word="rund"]	**rund**

8.5. Internetadressen

Adressen von Korpora

Sprache	Kürzel	Name	Adresse
Deutsch	**AAC**	Austrian Academy Corpus	http://www.aac.ac.at/
Deutsch	**BAS**	Bayerisches Archiv für Sprachsignale	http://www.phonetik.uni-muenchen.de/forschung/bay_arch_sprsig/
Deutsch	**ČNPK**	Česko-německý paralení korpus/ Das tschechisch-deutsche Parallelkorpus	https://ske.fi.muni.cz/auth/corpora/
Deutsch	**DDD**	Deutsch Diachron Digital	http://www.deutschdiachrondigital.de/
Deutsch	**DeReKo**	Das Deutsche Referenzkorpus	http://www1.ids-mannheim.de/kl/projekte/korpora/
Deutsch	**deTenTen**	German TenTen corpus	https://ske.fi.muni.cz/
Deutsch	**DeuCze**	Das deutsch-tschechische Parallelkorpus	http://www.deucze.germanistik.uni-wuerzburg.de/
Deutsch	**deWac**	Deutsches Webkorpus	http://wacky.sslmit.unibo.it/doku.php auch unter
Deutsch	**DGD**	Datenbank für gesprochenes Deutsch	http://dgd.ids-mannheim.de/
Deutsch	**DWDS**	Das Digitale Wörterbuch der deutschen Sprache	http://dwds.de/
Deutsch	**Falko**	Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache	https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko
Deutsch	**FnhdC**	Das Bonner Frühneuhochdeutschkorpus	http://www.korpora.org/Fnhd/
Deutsch	**GeWiss**	Korpus gesprochener Wissenschaftssprache	https://gewiss.uni-leipzig.de/
Deutsch	**HGC**	Huge German Corpus	http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.html
Deutsch	**CHTK**	Schweizer Text Korpus	http://chtk.unibas.ch/
Deutsch	**InterCorp**	InterCorp	http://www.korpus.cz/intercorp/
Deutsch	**KGSR**	Bochumer Korpus der gesprochenen Sprache im Ruhrgebiet	http://www.ruhr-uni-bochum.de/kgsr/
Deutsch	**Korpus C4**	Korpus C4	http://www.korpus-c4.org/
Deutsch	**MHDBDB**	Mittelhochdeutsche Begriffsdatenbank	http://mhdbdb.sbg.ac.at/
Deutsch	**MULTEXT-East**	Multilingual Text Tools and Corpora for Central and Eastern European Languages	http://nl.ijs.si/ME/
Deutsch	**OPUS**	The Open Parallel Corpus	http://opus.lingfil.uu.se/
Deutsch	**Wortschatz-Portal**	Korpusbasiertes Wortschatz-Portal der Universität Leipzig	http://wortschatz.uni-leipzig.de/

Sprache	Kürzel	Name	Adresse
Englisch	**BNC**	British National Corpus	http://www.natcorp.ox.ac.uk/
Englisch	**GloWbE**	Global Web-Based English	http://corpus2.byu.edu/glowbe/
Englisch	**ukWac**	Web corpus of British English (Web-Korpus des britischen Englisch)	http://wacky.sslmit.unibo.it/ auch über: http://korpus.cz/
Französisch	**CRFP**	Le Corpus de Référence du Français Parlé	http://sites.univ-provence.fr/delic/corpus/index.html
Französisch	**frWac**	Corpus électronique de français constitué à partir du web (französisches Web-Korpus)	http://wacky.sslmit.unibo.it/ auch über: http://korpus.cz/
Griechisch	**EThEC/ HNC**	Εθνικός Θησαυρός Ελληνικής Γλώσσας / Helenic National Corpus (Griechisches Nationalkorpus)	http://hnc.ilsp.gr/en/
Italienisch	**CORIS**	Corpus di italiano scritto / Corpus of written Italian	http://corpora.dslo.unibo.it
Italienisch	**itWac**	Web corpus di italiano (italiensches Web-Korpus)	http://wacky.sslmit.unibo.it/ auch über: http://korpus.cz/
Kroatisch	**HNK**	Hrvatski nacionalni korpus (Kroatisches Nationalkorpus)	http://www.hnk.ffzg.hr/
Polnisch	**NKJP**	Narodowy Korpus Języka Polskiego (Nationalkorpus der Polnischen Sprache)	http://nkjp.pl
Russisch	**NKRJ**	Национальный корпус русского языка (Nationales Korpus der russischen Sprache)	http://www.ruscorpora.ru/
Slowakisch	**SNK**	Slovenský národný korpus (Slowakisches Nationalkorpus)	http://korpus.juls.savba.sk.
Slowenisch	**FIDAPLUS**	Korpus slovenskega jezika (Korpus der slowenischen Sprache	http://www.fidaplus.net/
Spanisch	**Corpus del Español**	CORPUS DEL ESPAÑOL (Korpus des Spanischen)	http://www.corpusdelespanol.org/
Tschechisch	**ČNK**	Český národní korpus (Tschechisches Nationalkorpus)	http://korpus.cz/
Türkisch	**TUD/ TNC**	Türkçe Ulusal Derlemi / Turkish National Corpus (Türkisches Nationalkorpus)	http://www.tnc.org.tr
Ungarisch	**MNSz/ HuNC**	Magyar Nemzeti Szövegtár/ Hungarian National Corpus (Ungarisches Nationalkorpus)	http://corpus.nytud.hu/mnsz/

Adressen von anderen (korpusähnlichen) Instrumenten

Kürzel	Name	Adresse
ADABA	Österreichisches Aussprachewörterbuch/ Österreichische Aussprachedatenbank	http://adaba.at/
FORWO	FORWO, das Aussprachewörterbuch	http://de.forvo.com/
Linguee	Linguee	http://www.linguee.de/
ParZU	ParZu - The Zurich Dependency Parser for German	http://kitt.cl.uzh.ch/kitt/parzu/
TextSTAT	TextSTAT 2.9. (Mathias Hüning)	http://neon.niederlandistik.fu-berlin.de/textstat/

Schlusswort

Was können nun Sprachkorpora den Deutschlernenden bringen? Diese Frage stand als Impulsgedanke für diese Publikation und wurde hoffentlich zumindest teilweise beantwortet. Die Beschreibungen im Abschnitt IV und V können aufgrund der stetigen technischen und graphischen Entwicklung des jeweiligen Instruments bereits zum Zeitpunkt des Erscheinens dieses Buches veraltet sein. Vieles musste hier natürlich nur in Andeutungen bleiben, die entworfenen Forschungsfragen müssten in eigenständigen Studien tiefgreifender angegangen und detaillierter behandelt werden. Dies war jedoch nicht das Ziel dieser Arbeit. Vielmehr sollte die Breite der Nutzungsmöglichkeiten gezeigt werden, obwohl auch diese für DaF und DaZ viel breiter sind als der Platz in diesem Buch.

Die Nutzung von Korpora allgemein (nicht nur im DaF/DaZ) bedeutet für jede/-n selbstständige Arbeit mit der Sprache, Entdeckung neuer Erkenntnisse über die Sprache, über die Kultur dieser Sprache. Darüber hinaus bieten Korpora auch Einblicke in Ereignisse, die in der Gesellschaft passiert sind. Auch bekommt man nämlich oft Lust den ganzen Text zu lesen, aus dem der Beleg oder die Belege stammen.
Trotz dieser und anderer Argumente, die für die Nutzung von Korpora sprechen, ist unbestritten, dass die Arbeit mit Korpora (vor allem in der Anfangsphase) auch viel unnötig verlorene Zeit bedeuten kann. Es ist nur zu hoffen, dass mit diesem Buch der Zeit- und Arbeitsaufwand, der auf technische Probleme zurückzuführen ist (wie etwa das Suchen des richtigen Korpus, seiner Eigenschaften, der Formulierung der Abfrage, der Bedeutung eines Kürzels), verringert werden kann.

Auch muss man bei der Arbeit mit Korpora bedenken, dass jedes Korpus (auch das DeReKo mit seinen Milliarden von Textwörtern) lediglich einen „winzigen“ Ausschnitt aus allen Texten darstellt, die in der jeweiligen Sprache getätigt werden. Wenn ein Wort oder ein Phrasem in einem Korpus nicht zu finden ist, bedeutet dies nicht, dass es in der Sprache nicht existieren. Aber: die Statistik lügt nicht. Wenn in einem großen Korpus (mit über einer Mrd. Worte) ein Wort nicht vorkommt, heißt es, dass dieses Wort allgemein gesehen und im Vergleich zu anderen Wörtern viel seltener gebraucht wird. Dieses Wort ist dann einerseits für die allgemeine Kommunikation eher unwichtig, muss daher nicht im allgemeinen Sprachunterricht vermittelt werden. Andererseits kann es aber sein, dass dieses Wort in einem spezifischen Sprachgebrauch häufig vorkommt.

Die Differenzen im Sprachgebrauch sind am besten zu sehen, wenn man Daten aus zwei unterschiedlich aufgebauten Korpora nimmt: In den Tabellen im Kap. 7 sind „Durchschnittswerte“ aus mehreren Korpora angeführt um die Objektivität der Angaben zu erhöhen. Vergleicht man die ersten hundert häufigsten Substantive eines ausgewogenen Korpus (z.B. DeReWo 2007) mit denselben eines nicht ausgewogenen Korpus (z.B. InterCorp_de), stellt man viele Überschneidungen, aber auch Unterschiede fest. Noch deutlicher ist es zu sehen, wenn man die zwanzig häufigsten Substantive (ihre Lemmata) in zwei unterschiedlichen Texten vergleicht.

Im Roman *Malina* von Ingeborg Bachmann sind es (nach Häufigkeit): *Vater, Zeit, Leben, Herr, Tag, Frau, Hand, Haus, Mann, Nacht, Welt, Augen, Wasser, Gesicht, Menschen, Tür, Kopf, Namen, Wort, Mutter*, in Pippi Langstrumpf von Astrid Lindgren hingegen: *Mädchen, Kind, Pferd, Herr, Schule, Papa, Villa, Mama, Lehrerin, Damen, Hand, Tag, Hause, Weile, Leute, Seil, Kunterbunt, Kopf, Frau* und *Tisch*. Daneben gibt es, obwohl die beiden Texte nur

wenig gemeinsame Züge aufweisen, auch volle Übereinstimmungen in der Lexik: *Frau, Hand, Haus, Herr, Kopf, Tag*. Einige weitere Lexeme unterscheiden sich nur stilistisch: *Vater/ Papa, Mutter/ Mama*. Diese Lexeme findet man auch im Kap. 7.2 auf S. 188. Sie gehören unumstritten zum „harten" Kern des deutschen Wortschatzes.

Alle synsemantischen Wörter, Pronomina und sogar viele Verben sind fast ident.
Dieses Phänomen lässt sich mit der Tatsache vergleichen, dass jeder Mensch neben den allgemeinen sprachlichen Mitteln auch seine „eigene Sprache" verwendet. Es betrifft selbstverständlich nicht nur den individuellen Wortschatz, sondern auch sprachliche Strukturen und kommunikative Strategien, die bei jedem Sprecher und jeder Sprecherin unterschiedlich beliebt sind und deswegen auch unterschiedlich häufig realisiert werden. Aus den Korpusdaten sind jedoch diejenigen sprachlichen Elemente zu erkennen, die (fast) jeder verwendet und versteht. Deswegen sollten eben diese den Sprachlernenden vermittelt werden.

Die Statistik lügt zwar nicht, sie darf aber auch nicht überschätzt werden. Die Daten der Korpora entsprechen nur der Realität der Texte, die im Korpus gespeichert sind - nicht der allgemeinen Realität der Sprache, die sowieso wegen ihrer Komplexität nicht fassbar ist. Darüber hinaus weisen alle Korpora noch viele technische, aber auch „faktische" Mängel auf, die nicht in absehbarer Zeit behoben werden oder kaum jemals behoben werden können. Die Gründe sind zahlreich und spiegeln sich in den häufigsten Problemen wider, auf die man während der Korpusarbeit stößt. Sie lassen sich grundsätzlich in zwei markante Problemfelder zusammenfassen: Korpusmängel und Datenfehler. Deshalb muss jegliche Korpusarbeit mit dem Wissen ablaufen, dass Korpora mangelhaft im Aufbau und weitaus nicht frei von Fehlern jeglicher Art sind.

Ein großes Manko der Korpuslandschaft ist das Fehlen von wirklich repräsentativen Korpora der gesprochenen Sprache (dazu auch Heine 2008: 6). Wenn es diese in der (hoffentlich nahen) Zukunft auch geben sollte, wären sie zur Zeit ihrer Veröffentlichung bereits veraltet, denn die gesprochene Sprache ändert sich viel schneller als die geschriebene Sprache. Dazu kommt das bekannte Phänomen, dass man beim Schreiben automatisch eher auf die Norm achtet als beim Sprechen. Für die Objektivierung der Daten, aus denen normative Werke frei von (traditionell) puristischen Einflüssen und Eingriffen entstehen könnten, wären jedoch große Korpora der gesprochenen Sprache eine enorme Hilfe.
Parallele Korpora müssen (vermutlich langfristig) auf gesprochene Sprache völlig verzichten und müssen sich damit begnügen, was der aktuellen gesprochenen Sprache am nächsten kommt, nämlich gegenwärtige belletristische Werke in professioneller Übersetzung.

Der korpuslinguistische Grundsatz „je mehr Daten, desto schärfer das Bild über das untersuchte Phänomen" kann abschreckende Wirkung haben. Zu viele Daten sind nicht bewältigbar, kein Mensch kann hunderttausende Konkordanzen in einer akzeptablen Zeitspanne sortieren. Deswegen ermöglichen gute Korpora eine Zufallsauswahl der Belege. Problematisch sind jedoch homonyme Formen, die von automatischen Textanalysatoren (Tagger, Parser) falsch erkannt wurden. Diese Fehler sind sehr häufig. Am besten ist dies an solchen Wörtern zu erkennen, die an ähnlichen Stellen im Satz (satztopographisch gemeint) unterschiedliche Funktionen einnehmen. Diese sind für die automatischen Analysatoren besonders schwierig richtig zu erkennen, daher ist die Fehlerquote extrem hoch. Beispielsweise kann die Wortform *schon* als Temporaladverb, Abtönungspartikel, Antwortpartikel oder (umgangssprachlich) als Imperativ des Verbs *schonen* vorkommen:

> "Danke, und du **schon** dich auch!"
> Wenn du **schon** mal wegen irgendwas an der Grenze erwischt wurdest
> Wohl habe ich gestern **schon** dich gesehen und begrüßt. "
> **Schon** ganz schön teuer!

Diese Formen wurden im Korpus InterCorp_de und im DeReKo teilweise falsch erkannt. Hier muss man wieder der Statistik vertrauen und eine Zufallsauswahl durchführen. Aus den Konkordanzen der Zufallsauswahl muss man die prozentuelle Vertretung (hier der einzelnen Funktionen des Wortes *schon*[58]) berechnen und auf die Gesamtmenge der Konkordanzen übertragen. Wie viele Belege in die Zufallsauswahl einbezogen werden sollen, hängt sehr stark vom Ziel der Arbeit ab: Für Arbeiten, deren Aufgabe die erste Orientierung in der Problematik ist (Seminararbeiten, Forschungsprojekte), reichen oft bis zu hundert Belege. Kleinere wissenschaftliche Arbeiten sollten mit 500 bis 1.000 Belegen arbeiten, größere Arbeiten können erst mit mehreren (zehn)tausenden Belegen das gewählte Problem seriös beleuchten.

Die Fehlerquote der Annotation variiert sehr stark. In professionell erstellten Korpora werden der überwiegenden Mehrheit der Wörter sowohl ihre richtigen Grundformen als auch die richtigen morphosyntaktischen (oder syntaktisch-semantischen) Angaben zugewiesen. Die Fehlerquote liegt bei automatisch annotierten Korpora im Schnitt um 7 bis 10%. Daher ist es notwendig, die Belege zumindest stichpunktartig durchzusehen und die Richtigkeit zu überprüfen. Die Annotationsinstrumente werden stets verbessert, daher sinkt auch die Fehlerquote. Das Aussortieren von falschen Belegen ist sicherlich eine der Arbeiten, die ohne Korpusdaten nicht existieren würden. Für gewisse qualitative Untersuchungen bzw. auch für die Vermittlung einiger Elemente braucht man nicht viele Belege, daher auch keine Korpora. Hier schließt sich bereits der Kreis zu den Gedanken im Kap. 1.
Ein häufiger Störfaktor sind auch Tippfehler. Viele Texte werden immer noch eingescannt, die Texterkennungsprogramme (OCR) entziffern nicht immer alle Buchstaben einwandfrei, darüber hinaus werden Texte unter dem heutigen hohen Zeitdruck mit vielen Fehlern herausgegeben.

Grundsätzlich sind aber Korpora Instrumente, die die Sprache (oder Sprachen im Vergleich) realitätsnah und objektiv erfassen können. Elemente oder Strukturen, über die früher hinweg geschaut wurde, können endlich objektiv überprüft werden. Beispiele, dass normative Werke diesbezüglich Lücken aufweisen, wurden im Kap. 6 gebracht. So gesehen sind Korpora Instrumente, mit denen „im DaF-Unterricht ein gesundes Gleichgewicht zwischen der Präskription und Deskription zu erreichen [ist]“ (Stuyckens/ Brône 2009: 8). In Anbetracht der Tatsache, dass man sich von präskriptiven Werken emanzipieren kann, dass plurizentrische Sprachen differenziert untersucht und vermittelt werden können (vgl. auch Sorger 2013: 44), sind Korpora auch Instrumente für die Demokratisierung der Sprachbetrachtung und Sprachvermittlung.

Längst haben auch viele Wirtschaftsbereiche erkannt, dass Korpora bzw. einige Korpusinstrumente (v.a. die Kookkurrenz/Kollokationsanalyse) ein hilfreiches Mittel zur Aufdeckung komplexerer Zusammenhänge bieten. Es werden Analysen von Krankenberichten zur Erstellung neuer Diagnosen durchgeführt, Fehlerkomplexe in der Technik und Kommunikation erforscht oder kommerzielle Aufträge über verschriebene Medikamente vergeben.

Auch im neuesten Fach der angewandten Linguistik, in der forensischen Linguistik, spielen Korpusinstrumente zum Nachweis sprachlicher Besonderheiten der Täter eine wichtige Rolle.

[58] Die Abfrage nach der Wortform *schon* liefert über 45.000 Belege, im DeReKo, Archiv TAGGED-T über eine Mio. Treffer.

Diejenigen, die mit DaF und DaZ zu tun haben, wissen, wie unterschiedlich Lernende und ihre Bedürfnisse sind, wie schwierig es ist, ein passendes Lehrbuch für eine ganze Lernergruppe zu finden, noch dazu ein solches, das dem aktuellen Stand der Sprache, womöglich auch noch in der Region entsprechen würde. Korpora können Lehrbücher nicht ersetzen, sie können sie aber hervorragend ergänzen. Deshalb soll die Nutzung von Sprachkorpora für jede/-n Sprachlehrer/-in ein selbstverständlicher Teil der Arbeit werden. Noch mehr: die Grundkenntnisse der Korpusarbeit sollten auch an die Lerner/-innen weitergegeben werden.

Für die Erforschung der Sprachen, für den Vergleich von Sprachen und für eine effektive Vermittlung jeder Sprache sind Korpora ein Instrument, um das man nicht herumkommt. Ein Lehrer oder eine Sprachwissenschaftlerin ohne ein Sprachkorpus ist schlichtweg wie ein Pathologe ohne Leiche/ eine Bank ohne Tresor/ eine Chemielehrerin ohne Labor oder ein Gärtner ohne Dünger.

Literaturverzeichnis

AAC (2014): Austrian Academy Corpus. Wien: Österreichische Akademie der Wissenschaften. http://www.aac.ac.at/ [20.7.2014]

Abel, Andrea/ Zanin, Renata (Hg.) (2011): Korpora in Lehre und Forschung. Bozen: BU, Bozen Univ. Press.

ADABA (2005-2014): Österreichisches Aussprachewörterbuch/ Österreichische Aussprachedatenbank. Graz: Forschungsstelle österreichisches Deutsch. http://adaba.at/ [27.7.2014]

Aijmer, Karin/ Altenberg, Bengt (Hg.) (1991): English corpus linguistics: studies in honour of Jan Svartvik. London: Longman.

Baker, Paul/ Hardie, Andrew/ McEnery, Tony (2006): A glossary of corpus linguistics. Edinburgh: Edinburgh Univ. Press.

Barlow, Michael (2009): ParaConc. http://www.paraconc.com/ [17.9.2013]

Barnickel, Klaus-Dieter (1992): Falsche Freunde: ein vergleichendes Wörterbuch Deutsch - Englisch. Heidelberg: Groos.

Baroni, Marco/ Bernardini, Silvia/ Ferraresi, Adriano/ Zanchetta, Eros (2009): The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In: Language Resources and Evaluation 43 (3). Springer Science + Business Media. S. 209-226.

BAS (1995-2014): Bayerisches Archiv für Sprachsignale. München: Institut für Phonetik und Sprachverarbeitung (IPS), Ludwig Maximilians Universität München. http://www.phonetik.uni-muenchen.de/forschung/bay_arch_sprsig/ [12.5.2014]

Belica, Cyril (1995): Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache. http://corpora.ids-mannheim.de/ [22.5.2012]

Berman, Stephen (2014): Sitemap zur Website des Proseminars Korpuslinguistik, SS 2014. Bochum: Ruhr-Universität. http://homepage.rub.de/Stephen.Berman/Korpuslinguistik/sitemap.html [28.7.2014]

Bibel Online (2014): CID - christliche internet dienst GmbH. Berlin. http://www.bibel-online.net/ [2.3.2014]

Biber, Douglas/ Conrad, Susan/ Reppen, Randi (1998): Corpus linguistics: investigating language structure and use. Cambridge: University Press.

BNC (2010): British National Corpus. Oxford: University of Oxford. http://www.natcorp.ox.ac.uk/ [25.7.2014]

Bopp, Sebastian (2010): Einführung in die Korpuslinguistik mit DeReKo und COSMAS II. Augsburg: Universität Augsburg, Germanistik. https://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/mitarbeiter/stelspass/materialien_lehrveranstaltungen/korpuslinguistik_dereko_cosmas2_bopp.pdf [20.7.2014]

Braun, Sabine (2005): Corpora4Learning.net. [Englische Korpora im Fremdsprachenunterricht]. Guildford, Surrey: University of Surrey. http://www.corpora4learning.net/ [17.7.2014]

Breyer, Yvonne Alexandra (2011): Corpora in Language Teaching and Learning. Potential, Evaluation, Challenges. Frankfurt (M) u.a.: Peter Lang.

Bubenhofer, Noah (2006-2013): Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge. http://www.bubenhofer.com/korpuslinguistik/ [7.4.2014]

Bubenhofer, Noah/ Konopka, Marek/ Schneider, Roman (2014): Präliminarien einer Korpusgrammatik. Tübingen: Narr.

Budin, Gerhard (2011): Die Sprache im Wandel der Zeit. In: Thema. Das Forschungsmagazin der ÖAW, 10/2011. Wien: Österreichische Akademie der Wissenschaften. S. 16-17.

Carstensen, Kai-Uwe/ Ebert, Christian/ Ebert, Cornelia/ Jekat, Susanne/ Klabunde, Ralf/ Langer, Hagen (Hrgs.) (2010): Computerlinguistik und Sprachtechnologie: eine Einführung. Heidelberg: Spektrum Akad. Verl.

Čermák, František (2005): Jak využívat Český národní korpus. Praha: Nakladatelství Lidové noviny.

Čermák, František/ Klégr, Aleš (2004): Modality in Czech and English. In: International Journal of Corpus Linguistics 9:1. S. 83-95.

Chiarcos, Christian/ Erjavec, Tomaz (2011): OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In: Proceedings of the 5th Linguistic Annotation Workshop (LAW-V). Portland (Oregon). S. 11-20.

Chomsky, Noam (1986): Knowledge of language: its nature, origin, and use. New York: Praeger.

CHTK (2008-2012): Schweizer Text Korpus. Zürich: Schweizer Textkoprus/ Schweizerisches Idiotikon. http://www.schweizer-textkorpus.ch/index.php/de/ [25.7.2014]

CORIS (2011) Corpus di italiano scritto/ Corpus of written Italian. Bologna: Università di Bologna. http://corpora.dslo.unibo.it [27.6.2014]

Corpus del Español (2002-2014): Davies, Mark: Corpus del Español: 100 million words, 1200s-1900s. http://www.corpusdelespanol.org [21.3.2014]

Cosmas II (2012): Corpus Search, Management and Analysis System. Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/cosmas2/ [15.7.2014]

ČNK (2014): Czech Nation Corpus (and other available corpora). Praha: Univerzita Karlova, Ústav Českého národního korpusu. http://korpus.cz/ [26.7.2014]

ČNK-Tagset (2014): Hajič, Jan/ Cvrček, Václav: Internetová příručka ČNK. Morfologické značky (tagy). http://wiki.korpus.cz/doku.php/seznamy:tagy [27.7.2014]

ČNPK (2000-2005): Česko-německý paralení korpus/ Das tschechisch-deutsche Parallelkorpus. https://ske.fi.muni.cz/auth/corpora/ [15.7.2014]

CRFP (2014): Le Corpus de Référence du Français Parlé. http://sites.univ-provence.fr/delic/corpus/index.html [20.6.2014]

Davies, Mark (2014): corpus.byu.edu. [Korpora an der Brigham Young University]. Provo, Utah: Birgham Young University. http://corpus.byu.edu/corpora.asp [12.7.2014]

DDD (2011): Deutsch Diachron Digital. Jena: Friedrich-Schiller-Universität u.a. http://www.deutschdiachrondigital.de/ [5.6.2014]

DeReKo (2014): Deutsches Referenzkorpus. Mannheim: Institut für deutsche Sprache. http://www1.ids-mannheim.de/kl.html [20.7.2014]

DeReWo (2009): Korpusbasierte Wortlisten DeReWo. Mannheim: Institut für Deutsche Sprache. http://www.ids-mannheim.de/kl/derewo/ [15.4.2014]

deTenTen (2014): German TenTen corpus. Brno: Masarykova univerzita. https://ske.fi.muni.cz/ [15.7.2014]

deWac (2013): Deutsches Webkorpus. http://wacky.sslmit.unibo.it/doku.php [20.7.2014]

DGD (2014): Datenbank für gesprochenes Deutsch. Mannheim: Institut für Deutsche Sprache. http://dgd.ids-mannheim.de/ [15.7.2014]

Dovalil, Vít/ Káňa, Tomáš/ Pelouškova, Hana/ Zbytovský, Štěpán/ Vavřín, Martin (2013): Korpus intercorp_de Version 6 vom 8. 4. 2013. Praha: Ústav Českého národního korpusu. http://www.korpus.cz [29.7.2014]

Drosdowski, Günther (1985): Die Dudenredaktion. In: Wimmer, Rainer (Hg.): Sprachkultur. Sprache der Gegenwart: Schriften des Instituts für Deutsche Sprache in Mannheim (63). Düsseldorf: Schwann.

Duden - Grammatik (2005): Kunkel-Razum, Kathrin/ Münzberg, Franciska (Redaktion): Duden: die Grammatik. Unentbehrlich für richtiges Deutsch. Mannheim: Dudenverlag.

Duden - Universalwörterbuch (2006): Wermke, Matthias/ Kunkel-Razum, Kathrin/ Scholze-Stubenrecht, Werner (Hg.): Duden - Deutsches Universalwörterbuch, 6. Aufl. Mannheim: Bibliographisches Institut & F. A. Brockhaus AG. [CD-ROM]

Ďurčo, Peter (1994): Probleme der allgemeinen und kontrastiven Phraseologie (am Beispiel Deutsch und Slowakisch). Heidelberg: Groos.

Ďurčo, Peter/ Banášová, Monika/ Hanzlíčková, Astrid (2010): Feste Wortverbindungen im Kontrast. Trnava: Univerzita sv. Cyrila a Metoda.

Ďurčo, Peter/ Kathrin Steyer (2012): Ein korpusbasiertes Beschreibungsmodell für die elektronische Sprichwortlexikografie. SprichWort-Plattform. http://sprichwort-plattform.org/attach/Ergebnisse/SW_Modell_steyer_durco.pdf [22.7.2014]

Dürscheid, Christa/ Elspaß, Stephan/ Ziegler, Arne (2014): Variantengrammatik des Standarddeutschen. [Ein internationales Projekt der Universitäten Graz, Salzburg, Zürich.] http://www.variantengrammatik.net/ [20.7.2014]

DWB (1854-1961): Grimm, Jacob/ Grimm, Wilhelm: Deutsches Wörterbuch. http://woerterbuchnetz.de/DWB/ [15.5.2014]

DWDS (2014): Das Digitale Wörterbuch der deutschen Sprache. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. http://dwds.de/ [15.7.2014]

Engel, Ulrich (1988): Deutsche Grammatik. Heidelberg: Groos.

EThEG/ HNC (1999-2009): Εθνικός Θησαυρός Ελληνικής Γλώσσας/ Helenic National Corpus. Athena: [Institut für Sprache und Sprachverarbeitung] http://hnc.ilsp.gr/en/ [17.7.2014]

Facchinetti, Roberta (Hg.) (2007): Corpus Linguistics 25 Years on. Amsterdam u.a.: Rodopi.

Fachlexikon DaF/DaZ (2010): Barkowski, Hans/ Krumm, Hans-Jürgen (Hg.): Fachlexikon Deutsch als Fremd- und Zweitsprache. Tübingen/ Basel: Francke.

Falko (2014): Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache. Berlin: Humboldt-Universität. https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko [20.7.2014]

FIDAPLUS (2014): Korpus slovenskega jezika. Ljubljana: Univerza v Ljubljani. http://www.fidaplus.net/ [27.7.2014]

Fleischer, Wolfgang (1969): Wortbildung der deutschen Gegenwartssprache. Leipzig: VEB Bibliographisches Institut.

FnhdC (2007): Das Bonner Frühneuhochdeutschkorpus. http://www.korpora.org/Fnhd/ [12.6.2014]

Forum Deutsch als Fremdsprache (1996-2012): Internetservice für den Unterricht Deutsch als Fremdsprache. Düsseldorf: Institut für Internationale Kommunikation e.V. http://www.deutsch-als-fremdsprache.de/ [14.3.2014]

FORWO (2008-2014): FORWO, das Aussprachewörterbuch. http://de.forvo.com/ [27.7.2014]

Francis, Nelson W. (1992): Language Corpora B.C. In: Svartvik, Jan (Hg.): Directions in Corpus Linguistics. Berlin: de Gruyter. S. 17-32.

frWac (2013): Corpus électronique de français constitué à partir du web. http://wacky.sslmit.unibo.it/ [27.7.2014]

GeWiss (2014): Gesprochene Wissenschaftssprache. Leipzig: Universität Leipzig. https://gewiss.uni-leipzig.de/ [20.7.2014]

GloWbE (2013): Davies, Mark: Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. http://corpus2.byu.edu/glowbe/ [17.4.2014]

Granger, Sylviane/ Lerot, Jacques/ Petch-Tyson, Stephanie (2003): Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam u.a.: Rodopi.

Harries, Tony/ Moreno Jaén, María (Hg.) (2010): Corpus linguistics in language teaching. Bern/ Wien u.a.: Peter Lang.

Hausmann, Franz Josef (2003): Was sind eigentlich Kollokationen? In: Steyer, Kathrin (Hg.): Wortverbindungen - mehr oder weniger fest. Berlin u.a.: de Gruyter. S. 309-334.

Heine, Antje (2008): Zur Nutzbarkeit der gegenwärtig verfügbaren deutschen Korpora für die Lerenerlexikografie Deutsch als Fremdsprache. Anspruch und Wirklichkeit. In: Deutsch als Fremdsprache 45/1. Leipzig: Universität Leipzig.

Heringer, Hans Jürgen (2012): Chunking: Synonymik des Deutschen korpusbasiert. Tübingen: Narr.

HGC (2013): Huge German Corpus. Stuttgart: Universität Stuttgart. http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/hgc.html [2.3.2014]

Hirschová, Milada (2013): Pragmatika v češtině. Praha: Univerzita Karlova, Karolinum.

HNK (2005-2013): Hrvatski nacionalni korpus. Zagreb: Sveučilište u Zagrebu. http://www.hnk.ffzg.hr/ [17.7.2014]

Hüning, Matthias (2013): TextSTAT - Simples Text Analyse Tool. Berlin: Freie Universität. http://neon.niederlandistik.fu-berlin.de/textstat/ [22.11.2013]

InterCorp (2011, 2014). Korpus InterCorp. Praha: Univerzita Karlova. http://ucnk.ff.cuni.cz/intercorp/?req=page:info [25.7.2014]

itWac (2013): Web corpus di italiano. http://wacky.sslmit.unibo.it/ [17.7.2014]

Jelínek, Tomáš/ Petkevič, Vladimír/ Rosen, Alexandr/ Skoumalová, Hana (2012): Czech Treebanking Unlimited. Praha: Univerzita Karlova. http://utkl.ff.cuni.cz/synttb/Team_Czech_Treebanking_Unlimited.pdf [16.7.2014]

Káňa, Tomáš (2006): Korpuslinguistik - eine übersehene Herausforderung für den Deutschunterricht? In: Krumm, Hans-Jürgen (Hg.): Theorie und Praxis. Österreichische Beiträge zu Deutsch als Fremdsprache 9, 2005: Schwerpunkt: Innovationen - Neue Wege im Deutschunterricht. Innsbruck u.a.: StudienVerlag. S. 99-115.

Káňa, Tomáš (2006): Zur Problematik einiger tschechischer Ortsnamen in deutschen Texten. In: Lingua viva 2,1. České Budějovice: Jihočeská univerzita. S. 7-19.

Káňa, Tomáš (2007): Illokutive Kraft der Quellenangaben in Rundfunknachrichten. In: Balaskó, Maria/ Szatmári, Petra (Hg.): Sprach- und literaturwissenschaftliche Brückenschläge. München: Lincom.

Káňa, Tomáš (2008): Elektronische Sprachkorpora in Wissenschaft und Unterricht DaF/DaZ - einige Vorschläge für die Nutzung der elektronischen Instrumente. In: Krumm, Hans-Jürgen/ Portmann-Tselikas, Paul R. (Hg.): Theorie und Praxis: Österreichische Beiträge zu Deutsch als Fremdsprache. Schwerpunkt: Wortschatz (Serie A 11/2007). Innsbruck u.a.: StudienVerlag. S. 123-136.

Káňa, Tomáš (2010): Einige tschechische Flussnamen in elektronischen Korpora. In: Bock, Bettina (Hg.): Aspekte der Sprachwissenschaft. Hamburg: Verlag Dr. Kovač. S. 437-445.

Káňa, Tomáš (2012): Wortbildung: Umriss der Theorie mit Aufgaben und Übungen. Brno: Masarykova univerzita. https://is.muni.cz/elportal/?id=1071872 [15.6.2014]

Káňa, Tomáš / Hana Pelouškova (2009): Deutsch und Tschechisch im Vergleich. Korpusbasierte linguistische Studien. Brno: Masarykova univerzita.

Káňa, Tomáš / Hana Pelouškova (2011): Deutsch und Tschechisch im Vergleich: Korpusbasierte linguistische Studien II. Brno: Masarykova univerzita.

Káňa, Tomáš / Pelouškova, Hana (2006): Elektronische Korpora in Tschechien und das tschechisch-deutsche Parallelkorpus. In: Kettemann, Bernhard/ Marko, Georg (Hg.) Planing, Gluing and Painting Corpora. Frankfurt (M): Lang. S. 27-46.

Káňa, Tomáš/ Pelouškova, Hana (2010): Česko-německý paralelní korpus. Brno: Masarykova univerzita. http://www.ped.muni.cz/katedry-a-instituty/nemecky-jazyk-literatura/aktivity/cesko-nemecky-paralelni-korpus/ [20.7.2014]

Káňa, Tomáš/ Vavřín, Martin (2011): Das Korpus InterCorp (Deutsche Fassung). http://ucnk.ff.cuni.cz/intercorp/?lang=de [27.7.2014]

Káňa, Tomáš: (2010): Synthetische substantivische Diminutive im Tschechischen und ihre strukturellen Entsprechungen im Deutschen. In: Kratochvílová, Iva/ Wolf, Norbert Richard (Hg.): Kompendium Korpuslinguistik: Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. Heidelberg: Universitätsverlag Winter. S. 235-242.

KGSR (2014): Bochumer Korpus der gesprochenen Sprache im Ruhrgebiet. Bochum: Ruhr-Universität. http://www.ruhr-uni-bochum.de/kgsr/ [17.7.2014]

Kilgarriff, Adam/ Husák, Miloš /McAdam, Katy/ Rundell, Michael/ Rychlý, Pavel (2008): GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings

of the XIII EURALEX International Congress. Barcelona: Institut Universitari de Lingüística Aplicada. S. 425-432.

Kilgarriff, Adam/ Baisa, Vít/, Bušta, Jan / Jakubíček, Miloš/ Kovář, Vojtěch/ Michelfeit, Jan/ Rychlý, Pavel/ Suchomel, Vít (2014): The Sketch Engine: ten years on. In: Lexicography ASIALEX (2014) 1. Berlin/ Heidelberg: Springer. S 7-36.

Kluge, Friedrich (2002): Etymologisches Wörterbuch. Berlin u.a.: de Gruyter.

Knittlová, Dagmar/ Grygová, Bronislava/ Zehnalová, Jitka (2010): Překlad a překládání. Olomouc: Univerzita Palackého v Olomouci.

Kocek, Jan/ Kopřivová, Marie/ Kučera (Hg.) (2000): Český národní korpus: Úvod a příručka uživatele. Praha: Univerzita Karlova, Ústav Českého národního korpusu.

Konopka, Marek/ Kubczak, Jacqueline/ Mair, Christian/ Štícha, František/ Waßner, Ulrich Hermann (Hg.) (2011): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.-24.9.2009. Tübingen: Narr.

Korpus-C4 (2014): DWDS - AAC - Korpus Südtirol - Schweizer Textkorpus. http://www.korpus-c4.org/ [12.7.2014]

Kotulková, Veronika/ Kratochvílová, Iva/ Rykalová, Gabriela/ Wolf, Norbert Richard (2005-2010): DeuCze: Das deutsch-tschechische Parallelkorpus. Würzburg/ Opava: http://www.deucze.germanistik.uni-wuerzburg.de/ [16.6.2014]

Krajka, Jarosław (2007): Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections. CORELL: Computer Resources for Language Learning 1. S 36-55. http://www.ucam.edu/sites/default/files/corell/JKrajka.pdf [15.7.2014]

Kundera, Milan (1970): Scherz. München: Deutscher Taschenbuch Verlag.

Kurier (2014): Infokampagne zur MaHü: Experte glaubt nicht an Korruption. http://kurier.at/chronik/wien/infokampagne-zur-mariahilfer-strasse-experte-sickinger-glaubt-nicht-an-korruption/44.661.945 [07.01.2014]

Lawler, John M./ Aristar-Dry, Helen (Hg.) (1998): Using computers in linguistics: a practical guide. London u.a.: Routledge.

Leech, Geoffrey (1991): The State of the Art in Corpus Linguistics. In: Aijmer, Karin/ Altenberg, Bengt (Hg.): English Corpus Linguistics. London: Longman. S. 8-29.

Leech, Geoffrey (1997): Teaching and Language Corpora: a Convergence. In: Wichmann, Anne/ Fligelstone, Steven/ McEnery, Tony/ Knowles, Gerry (Hg.): Teaching and language Corpora. London u.a.: Longman.

Lemnitzer, Lothar/ Zinsmeister, Heike (2010): Korpuslinguistik: eine Einführung. Tübingen: Narr.

Lewandowski, Theodor (1994): Linguistisches Wörterbuch. Heidelberg u.a.: Quelle & Meyer.

Lewis, Michael (1996): The lexical approach: the state of ELT and a way forward. Hove: Language Teaching Publ.

Linguee (2013): Linguee. Köln: Linguee GmbH. http://www.linguee.de/ [17.7.2014]

Lišková, Danuša (2010): Peniaze - banky - burzy. In: Medvecká, Ľubica/ Šoltys, Jaroslav (Hg.): Odborný preklad 5: terminológia bankovníctva a finančníctva v súvislosti s prechodom na euro. Studia Translatologica Bratislavensia 20. Bratislava: AnaPress. S. 22-33.

Lüdeling, Anke/ Kytö, Merja (Hg.) (2008): Corpus Linguistics. An International Handbook, vol. 1. Berlin u.a.: de Gruyter.

Lüdeling, Anke/ Walter, Maik (2010): Korpuslinguistik. In: Fandrych, Christian/ Hufeisen, Britta/ Krumm, Hans-Jürgen/ Riemer, Claudia (Hg.): Deutsch als Fremdsprache und Zweitsprache. Ein internationales Handbuch. Berlin u.a.: de Gruyter. S. 315-322.

Mair, Rebecca (2013): buk vs. backte; buken vs. backten; gebackt vs. gebacken [unveröffentlichte Seminararbeit]. Wien: Universität Wien.

McEnery, Tony/ Wilson, Andrew (2001): Corpus linguistics: an introduction. Edinburgh: Univ. Press.

Merkel, Silke/ Schmidt, Thomas (2009): Korpora gesprochener Sprache im Netz - eine Umschau. In: Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion, 10 (2009). http://www.gespraechsforschung-ozs.de/heft2009/heft2009.html [12.2.2013]

Metzler (2000): Glück, Helmut: Metzler Lexikon Sprache. Stuttgart u.a.: Metzler.

MHDBDB (2012): Mittelhochdeutsche Begriffsdatenbank. Salzburg: Universität Salzburg. http://mhdbdb.sbg.ac.at/ [27.7.2014]

MNSz (1998-2006): Magyar Nemzeti Szövegtár/ Hungarian National Corpus. Budapest: Magyar tudományos akadémia. http://corpus.nytud.hu/mnsz/ [27.7.2014]

Moudraia, Olga (2001): Lexical Approach to Second Language Teaching. Washington: Center for Applied Linguistics. (ERIC Digest, EDO-FL-01-02).

MULTEXT-East (2013): Multilingual Text Tools and Corpora for Central and Eastern European Languages. http://nl.ijs.si/ME/ [12.7.2014]

NKJP (2008-2012): Narodowy Korpus Języka Polskiego. http://nkjp.pl [27.7.2014]

NKRJa (2003-2014): Национальный корпус русского языка http://www.ruscorpora.ru/ [14.6.2014]

OPUS (2014): The Open Parallel Corpus. Uppsala: Uppsala Universitet. http://opus.lingfil.uu.se/ [12.6.2014]

Österreichisches Wörterbuch (2001): Back, Otto/ Benedikt, Erich/ Blümel, Karl/ Ebner, Jakob/ Hornung, Maria/ Möcker, Hermann/ Pohl, Heinz-Dieter/ Tatzreiter, Herbert: Österreichisches Wörterbuch. Wien: ÖBV & HPT VerlagsGmbH.

ParZu (2014): ParZu - The Zurich Dependency Parser for German. Zürich: Universität Zürich. http://kitt.cl.uzh.ch/kitt/parzu/ [20.7.2014]

Pelouškova, Hana (2013): Das Projekt InterCorp und seine Rolle in der Deutschlehrerausbildung und Forschung. In: Slowakische Zeitschrift für Germanistik. Banská Bystrica: Verband der Deutschlehrer und Germanisten der Slowakei.

Perkuhn, Rainer/ Belica, Cyril (2004): Eine kurze Einführung in die Kookkurrenzanalyse und syntagmatische Muster. Mannheim: Institut für Deutsche Sprache. http://www1.ids-mannheim.de/kl/misc/tutorial.html [22.6.2014]

PONS Deutsch - Englisch (1998): Breitsprecher, Roland/ Terrell, Peter/ Schnorr, Veronika/ Morris, Wendy V.A: PONS Globalwörterbuch Deutsch - Englisch. Wien: ÖBV.

profil (2009): Interview "Hausbesetzer sind konservativ": Blixa Bargeld im Interview mit profil. http://www.profil.at/articles/0910/560/235752/hausbesetzer-blixa-bargeld-interview [14.6.2014]

Reznicek, Marc/ Lüdeling, Anke/ Krummes, Cedric/ Schwantuschke, Franziska/ Walter, Maik/ Schmidt, Karin/ Hirschmann, Hagen/ Andreas, Torsten (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Berlin: Humboldt-Universität zu Berlin. http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01. [2.5.2014]

RFTagger - German. Brighton: Lexical Computing Ltd. http://www.sketchengine.co.uk/documentation/wiki/tagsets/german_rftagger [22.7.2014]

Römer, Ute/ Schulze, Rainer (Hg.) (2009): Exploring the lexis-grammar interface. Amsterdam u.a.: Benjamins.

Rosen, Alexandr (2012): Grammar-based treebank a happy marriage of empiricism and theory? Praha: Univerzita Karlova. http://utkl.ff.cuni.cz/synttb/Rosen_HappyMarriage.pdf [27.7.2014]

Rösch, Heidi (2011): Deutsch als Zweit- und Fremdsprache. Berlin: Akad.-Verl.

Rychlý, Pavel (2008): A Lexicographer-Friendly Association Score. In RASLAN 2008. Brno: Masarykova Univerzita. S. 6-9.

Sennrich, Rico/ Schneider, Gerold/ Volk, Martin/ Warin, Martin (2009): A New Hybrid Dependency Parser for German. In: Chiarcos, Christian/ de Castilho, Richard Eckart/ Stede, Manfred: Von der Form zur Bedeutung: Texte automatisch verarbeiten/ From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009. Tübingen. S. 115-124.

Schiller, Anne/ Teufel, Simone/ Stöckert, Christine (1995): Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. Stuttgart: Universität Stuttgart. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/stts_guide.pdf [18.5.2014]

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger2.pdf [20.11.2013]

Schmid, Helmut/ Laws, Florian (2008): Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. Manchester: COLING 2008. http://www.cis.uni-muenchen.de/~schmid/papers/Schmid-Laws.pdf [23.3.2014]

Sinclair, John McHardy (Hg.) (2004): How to use corpora in language teaching. Amsterdam u.a.: Benjamins.

Skalička, Vladimír (1957): Vztah morfologie a syntaxe. In: Slovo a slovesnost, Jg. 18 /1957, Nr. 2. S. 65-71. http://sas.ujc.cas.cz/archiv.php?art=885 [12.4.2014]

Sketch Engine (2014): Sketch Engine. Text corpora and corpus tools for all. Brighton: Lexical Computing Ltd. http://www.sketchengine.co.uk/ [22.7.2014]

SNK (2013): Slovenský národný korpus. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. http://korpus.juls.savba.sk [27.7.2014]

Sorger, Brigitte (2012): Der Internationale Deutschlehrerverband und seine Sprachenpolitik - Ein Beitrag zur Fachgeschichte von Deutsch als Fremdsprache. Innsbruck: Studienverlag.

Sorger, Brigitte (2013): Institutions- und sprachenpolitische Aspekte des DACH-Konzepts. In: Demmig, Silvia/ Hägi, Sara/ Schweiger, Hannes (Hg.): DACH-Landeskunde. Theorie - Geschichte - Praxis. München: Iudicium. S. 32-48.

Sorger, Brigitte/ Káňa, Tomáš / Janíková, Věra/ Reitbrecht, Sandra/Brychová, Alice (2013): Schreiben in mehreren Sprachen. Deutsch nach Englisch: Mehrsprachigkeit und ihr Einfluss auf die Textkompetenz. Brno: Tribun EU.

Sparling, Don (1989): English or Czenglish? Praha: Státní pedagogické nakladatelství.

Steinbügl, Birgit (2005): Deutsch-englische Kollokationen: Erfassung in zweisprachigen Wörterbüchern und Grenzen der korpusbasierten Analyse. Tübingen: Max Niemeyer.

Storjohann, Petra (2005). Corpus-driven vs. corpus-based approach to the study of relational patterns. Proceedings of the Corpus Linguistics conference 2005 in Birmingham. http://www.corpus.bham.ac.uk/conference2005/index.htm [4.3.2014]

STTS Tag Table (1995/1999): Stuttgart - Tübinger Tag Set. Stuttgart: Universität Stuttgart. http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html [20.7.2014]

Stuyckens, Geert/ Brône, Geert (2009): Brauchbarkeit von Korpora des geschriebenen Deutsch für DaF-Lehrende. Eine Fallstudie. In: Deutsch als Fremdsprache. Jg. 46, 2009/1. Leipzig: Universität Leipzig.

Swan, Michael (1995): Practical English usage. Oxford: University Press.

Swan, Michael (2006): Chunks in the Classroom: Let´s Not Go Overboard. The Teacher Trainer 20/3, 2006. http://www.mikeswan.co.uk/elt-applied-linguistics/chunks-in-the-classroom.htm [20.5.2014]

Tadić, Marko (2009): New version of the Croatian National Corpus. In: Hlaváčková, Dana/ Horák, Aleš/ Osolsobě, Klára/ Rychlý, Pavel (Hg.): After Half a Century of Slavonic Natural Language Processing. Brno: Masaryk University. S. 199-205.

TenTen corpora (2004-2014): Kilgarriff, Adam/ Rychlý, Pavel/ Smrž, Pavel/ Tugwell, David: The Sketch Engine. Proc EURALEX 2004. France: Lorient. S. 105-116. http://www.sketchengine.co.uk. [27.7.2014]

TextSTAT (2014): Hüning, Mathias: TextSTAT 2.9. http://neon.niederlandistik.fu-berlin.de/textstat/ [12.6.2014]

Tiedemann, Jörg (2007): Building a Multilingual Parallel Subtitle Corpus. In Proceedings of CLIN 17. Belgium: Leuven, Alfa Informatica, University of Groningen. http://stp.lingfil.uu.se/~joerg/paper/clin17.pdf [27.7.2014]

Tiedemann, Jörg (2012): Parallel Data, Tools and Interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). S. 2214-2218

Tiedemann, Jörg (2009): News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: Nicolov, Nicolas/ Bontcheva, Kalina/ Angelova, Galina/ Mitkov, Ruslan (Hg.): Recent Advances in Natural Language Processing (vol. V). Amsterdam u.a.: Benjamins. S. 237-248.

TNC (2008-2012): Türkçe Ulusal Derlemi/ Turkish National Corpus. Mersin: Mersin Üniversitesi. http://www.tnc.org.tr/ [5.5.2014]

Tschirner, Erwin (2005): Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Heine, Antje/ Henning, Mathilde/ Tschirner, Erwin: Deutsch als Fremdsprache - Konturen und Perspektiven eines Faches. Festschrift für Barbara Wotjak zum 65. Geburtstag. München: Iudicum.

ukWac (2013): Web corpus of British English. http://wacky.sslmit.unibo.it/doku.php [27.7.2014]

Uri, Helene (2008): Nur die Stärksten überleben. München: Piper.

Variantenwörterbuch (2004): Ulrich Ammon/ Bickel, Hans/ Ebner, Jakob: Variantenwörterbuch des Deutschen: die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin u.a.: de Gruyter.

Westhof, Gerard J. (1987): Didaktik des Leseverstehens. Strategien des voraussagenden Lesens mit Übungsprogrammen. Ismaning: Hueber.

Westhof, Gerard J. (1991): Leseverstehen: Lesen, Lernen, Lehren. In: Andenmatten, Sigrid/ Bruder, Otto/ Faucherre, Alain/ Langer, Michael/ Schwarz, Alexander (Hg.): Verstehen im Deutschunterricht: Didaktik des Lese- und Hörverstehens im Fach Deutsch als Fremdsprache. Bulletin CILA (Schweizerische Hochschulkomission für angewandte Sprachwissenschaft), 1991/ 53. Neuchâtel: Université de Neuchâtel.

Wichmann, Anne/ Fligelstone, Steven/ McEnery, Tony/ Knowles, Gerry (Hg.) (1997): Teaching and language corpora. London u.a.: Longman.

Wikipedia: Übersetzungen (2014): Koordination von Übersetzungen aus anderen Sprachversionen der Wikipedia in die deutschsprachige. Absatz: Kopieren der Versionsgeschichte. http://de.wikipedia.org/wiki/Wikipedia:%C3%9Cbersetzungen#Lizenzfragen:_Urheberrecht_und_Originaltext. [13.1.14]

Wilson, Andrew/ Archer, Dawn/ Rayson, Paul (Hg.) (2006): Corpus linguistics around the world. Amsterdam u.a.: Rodopi.

Wortschatz-Portal (1998-2014): Deutscher Wortschatz. Leipzig: Universität Leipzig. http://wortschatz.uni-leipzig.de/ [15.7.2014]

Wray, Alison (2002): Formulaic Language and the Lexicon. Cambridge: University Press.

Zapletal, Štěpán/ Jungwirth, Karel/ Kouřimská, Milada (1980): Praktická mluvnice němčiny. Praha: Státní pedagogické nakladatelství.

Zifonun, Gisela/ Hoffmann, Ludger/ Strecker, Bruno/ Ballweg, Joachim/ Braußе, Ursula/ Breindl, Eva/ Engel, Ulrich/ Frosch, Helmut/ Hoberg, Ursula/ Vorderwülbecke, Klaus (1997): Grammatik der deutschen Sprache. Berlin u.a.: de Gruyter.

Index

Richtungsangabe **128**
sprachliche Angabe der Richtung, realisiert durch Lokalpräpositionen

Segment **20**
(allgemein) ein Teil eines Textes (Absatz, Satz…); (im Parallelkorpus) minimale Einheit, die aligniert (→ siehe Alignment) wird.

SNK **42**
Slovenský národný korpus (Slowakisches Nationalkorpus)

Sprachkorpus **13f**
elektronische Textdatenbank mit Suchmöglichkeiten nach sprachlichen Elementen

STTS **204**
Stuttgart - Tübingen Tagset: Verzeichnis der morphosyntaktischen Zeichen (u.a. für Deutsch) in annotierten Korpora

Suchfeldeingabe → siehe Eingabe **15**

Suffix **144, 203**
gebundenes Wortelement (existiert in der Sprache nicht frei) positioniert nach der Wortbasis

synchrones Korpus **24**
Korpus mit gegenwärtigen Texten

syntagmatische Muster **69, 166**
Ergebnis einer Kookkurrenzanalyse (DeReKo) in Form konkreter Verbindungen (Chunks)

Tag **19, 204**
morphosyntaktisches Zeichen/ Angabe über morphosyntaktische Funktion des Tokens

Tagger **19**
Programm für die morphosyntaktische Annotation von Korpustexten

Tagset **88, 204ff**
Liste der morphosyntaktischen Symbole und Zeichen

TextSTAT **105ff**
ein frei zugängliches Konkordanzprogramm

Token **29, 72**
formal selbstständige Einheit (Zeichenkette oder ein Zeichen) im Korpustext

Tokenisierung **29**
Segmentierung der Texte in kleinere Einheiten: Absätze, Sätze, Wörter und Satzzeichen (in allen gängigen Korpora); in geparsten Korpora: Segmentierung in Phrasen und Satzkonstituente

Treebank **20, 114**
Korpora mit geparsten (→ siehe Parser) Texten

TreeTagger **88, 206ff**
Korpustool für Annotationen einzelner Wörter mit morphosyntaktischen Informationen inkl. Zuweisung der Grundformen

Sprachkorpora
in Unterricht und Forschung DaF/DaZ

Tomáš Káňa, Ph.D., Mgr.
Herausgegeben von der Masaryk-Universität 2014
Auflagenhöhe: 100 Exemplare
1. Auflage, 2014
MSD, spol. s r.o., Lidická 23, 602 00 Brno, www.msdbrno.cz

ISBN 978-80-210-6994-7