Univerzita Karlova

Filozofická fakulta

Ústav germánských studií

Diplomová práce

Bc. Jitka Křesťanová

Vergleich der Ergebnisse der Kookkurrenzdatenbank (CCDB) und der Kookkurrenzanalyse

Porovnání výsledků kookurenční databanky (CCDB) a kookurenční analýzy

Comparing results of the co-occurrence database (CCDB) and the co-occurrence analysis

Praha 2017 vedoucí práce: Mgr. Věra Hejhalová, Ph.D.

Na tomto místě bych ráda poděkovala především vedoucí své diplomové práce, Mgr. Věře Hejhalové, Ph.D., za všestrannou pomoc a podnětné návrhy ke zpracování této práce.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze, dne 21. července 2017

……

Jméno a příjmení

Klíčová slova:

CCDB, kookurenční databanka, kookurenční analýza, korpus, jazykový korpus, korpusová lingvistika, IDS Mannheim, DeReKo, COSMAS II, analýza

Keywords:

CCDB, co-occurrence database, co-occurrence analysis, corpus, corpus linguistics, IDS Mannheim, DeReKo, COSMAS II, analysis

Schlüsselwörter:

CCDB, Kookkurrenzdatenbank, Kookkurrenzanalyse, Korpus, Sprachkorpus, Korpuslinguistik, IDS Mannheim, DeReKo, COSMAS II, Analyse

Abstrakt:

Předkládaná práce se zabývá korpusovou lingvistikou a blíže popisuje dvě aplikace, které zpracovávají data z korpusu DeReKo přístupem corpus-driven. Jedná se o kookurenční analýzu a Kookurenční databanku. Cílem práce je jednak zhodnotit, zda se výsledky získané kookurenční analýzou současného korpusu DeReKo liší od výsledků Kookurenční databanky, která byla vytvořena na korpusu menšího rozsahu. Kromě toho práce nabízí názorné příklady využití obou aplikací a zhodnocení jejich efektivnosti v závislosti na cíli výzkumu. Teoretická část práce pojednává o terminologii korpusové lingvistiky a o zmíněných korpusech, které jsou podkladem pro praktickou část práce. Empirickou část práce tvoří analýzy náhodně vybraných slov (jedno od každého slovního druhu) v obou aplikacích. Výsledky potvrzují, že data získaná pomocí Kookurenční databanky a kookurenční analýzy jsou v mnoha ohledech odlišná a potvrzují tak hypotézu, že velikost korpusu hraje ve výsledcích zásadní roli. Obě aplikace mají svá pozitiva i negativa. Práce nabízí jejich ucelený přehled a poskytuje tak uživateli návod, jak s oběma aplikacemi pracovat co nejefektivněji.

Abstract:

This paper deals with corpus linguistics. There are two applications under its scrutiny. Both of these applications are processing data from the corpus DeReKo via corpus-driven approach. It is a co-occurrence analysis and a Co-occurrence database. The aim of the work is to evaluate whether the results obtained by the co-occurrence analysis of the current scope of DeReKo are different from the results of the Co-occurrence database, which was created on a basis of a smaller scale corpus. In addition, this thesis offers illustrative examples of the use of both applications and the evaluation of their effectiveness, depending on the purpose of the research. The theoretical part of the thesis deals with the terminology of corpus linguistics and with the mentioned corpuses, which serve as a basis for the practical part of the thesis. The empirical part of the thesis consists of analyses of the randomly picked words (one from each word class) in both applications. The results confirm that the data obtained with Co-occurrence database and co-occurrence analysis are in many respects different and thus confirm the hypothesis that the corpus size plays a crucial role in the results. Both

applications have their advantages and disadvantages. The paper offers a comprehensive overview and by doing so it provides the user with instructions how to work with both applications in the most effective way.

Abstract:

Die vorliegende Arbeit beschäftigt sich mit der Korpuslinguistik und beschreibt näher zwei Applikationen, die die Daten vom Korpus DeReKo bearbeiten und die sich auf den Corpus-Driven-Zugang stützen. Es handelt sich um die Kookkurrenzanalyse (KA) und die Kookkurrenzdatenbank (CCDB). Das Ziel der Arbeit ist zu bewerten, ob sich die Ergebnisse, die im Rahmen der Kookkurrenzanalyse des gegenwärtigen Korpus DeReKo gewonnen wurden, von den Ergebnissen der Kookkurrenzdatenbank, die auf einem kleineren Korpusumfang basieren, voneinander unterscheiden. Daneben bietet die Arbeit anschauliche Beispiele von Verwendung beider Applikationen und bewertet ihre Effektivität in Abhängigkeit vom Forschungsziel. Der theoretische Teil behandelt die Terminologie der Korpuslinguistik und die Textkorpora, die als Grundlage für den empirischen Teil dienen. Den empirischen Teil der Arbeit bilden die Analysen zufällig ausgewählter Wörter (je eine Wortart) in beiden Applikationen. Die Ergebnisse bestätigen, dass die CCDB- und KA-Daten in vieler Hinsicht unterschiedlich sind. Damit wird die Hypothese bestätigt, dass die Korpusgröße eine grundlegende Rolle in den Ergebnissen spielt. Beide Applikationen weisen sowohl Vorteile als auch Nachteile auf. Die Arbeit bietet ihren komplexen Überblick und auch eine Anleitung, wie man mit beiden Applikationen auf möglichst effiziente Weise arbeiten kann.

Inhalt

1 Einleitung

Die vorliegende Diplomarbeit beschäftigt sich dem Vergleich der Ergebnisse der **Kookkurrenzdatenbank** (CCDB) und der **Kookkurrenzanalyse** (KA). Diese Applikationen, die die Daten vom Korpus DeReKo bearbeiten und die sich auf den Corpus-Driven-Zugang stützen, werden im Rahmen der Arbeit detailliert beschrieben.

Die Arbeit besteht sowohl aus einem theoretischen als auch aus zwei empirischen Teilen. Im theoretischen Teil werden zunächst die wichtigsten Termini aus dem Bereich der Korpuslinguistik definiert und die Mannheimer Korpora und deren Such- sowie Analysesysteme vorgestellt.

Der zweite Teil der Diplomarbeit beschreibt detailliert die beiden Analysenmethoden (KA und CCDB), ihre Anwendung wird mithilfe von Abbildungen belegt. In diesem Teil der Arbeit soll gezeigt werden, wie sich die Analysen durchführen lassen, wie sich die beiden Ergebnistabellen (KA und CCDB) auf den ersten Blick voneinander unterscheiden und wie man sie dechiffrieren soll. Die Parameter, die in der Einstellung der KA zur Verfügung stehen (Granularität, Autofokus, Lemmatisierung, Clusterzuordnung usw.), werden zunächst charakterisiert, folgend untereinander verglichen und ihre Vorteile (bzw. Nachteile) diskutiert. Die Ergebnisliste, die die CCDB liefert, wird in Form der Abbildung gezeigt und detailliert beschrieben. Dieser Teil der Arbeit bietet außerdem auch eine Anleitung, wie man mit beiden Applikationen auf möglichst effiziente Weise arbeiten kann. Im Rahmen der Arbeit soll auch darauf hingewiesen werden, wann für den Nutzer die CCDB und wann die KA günstiger sein kann.

Der letze Teil der Diplomarbeit widmet sich der praktischen und anschaulichen Durchführung der Analysen. Im Rahmen der Analysen wurden zehn zufällig ausgewählte Wörter je eine Wortart nominiert. Es werden das Substantiv *Gefahr*, das Adjektiv *schwarz*, der Artikel *der*, das Pronomen *man*, das Numerale *zweierlei*, das Verb *bestehen*, das Adverb *damals*, die Präposition *angesichts*, die Konjunktion *wenngleich* und die Interjektion *ach* in die Analyse einbezogen. Diese Analyse wird sowohl im Rahmen der kontinuierlich erweiterten KA des Instituts der Deutschen

Sprache (IDS) in Mannheim, als auch in der statischen Datenbank CCDB durchgeführt. Die Ergebnisse, die diese Analysen liefern, werden mithilfe von den Abbildungen beschrieben und verglichen. Der Vergleich der Ergebnisse der KA und der CCDB dient als Grundlage zur Erörterung der Differenzen und Gemeinsamkeiten zwischen den beiden Analysenmethoden.

Die Verfasserin dieser Abreit ist der Meinung, dass die Textmenge, die in einem Korpus enthalten ist, auch einen Einfluss auf die Ergebnisse der Kookkurrenzanalysen hat. Aus diesem Grund setzt die Verfasserin voraus, dass die Ergebnisse der CCDB und der KA Differenzen aufweisen können. Im Rahmen dieser Arbeit soll überprüft werden, ob die kontinuierliche Erweiterung des Korpus die Ergebnisse der Analyse beeinflusst und ändert. Dank dieser Untersuchung soll die Hypothese der Verfasserin bestätigt oder widerlegt werden. Im Rahmen der vorliegenden Arbeit werden auch die Vor- und Nachteile (z.B. die Überschaubarkeit und Benutzerfreundlichkeit) der beiden Analysenmethoden erwähnt und zusammengefasst.

2 Der theoretische Teil

Um die Kookkurrenzanalyse und die Kookkurrenzdatenbank besser verstehen zu können, müssen am Anfang dieser Arbeit die Grundbegriffe aus dem Bereich der Korpuslinguistik erklärt werden.

2.1 Korpuslinguistik und Textkorpora

Die **Korpuslinguistik** ist ein Bereich der Sprachwissenschaft und untersucht die Sprache in ihrem Gebrauch. Sie setzt sich zum Ziel, neue Erkenntnisse (über Sprache generell oder über bestimmte einzelne Sprachen) zu erlangen, neue Einsichten in die Strukturen, Gesetzmäßigkeiten, Eigenschaften und Funktionen von Sprache zu gewinnen oder bestehende Hypothesen zu überprüfen (bestätigen oder widerlegen). Als Grundlage dafür dienen die quantitativen oder qualitativen Daten. Diese Daten gewinnt man aus der Analyse von Korpora geschriebener oder gesprochener Sprache. Die Aufgabe der Korpuslinguistik ist die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind (vgl. LEMNITZER, ZINSMEISTER, 2006, 9). Korpuslinguistik ist eine wissenschaftliche Disziplin, d.h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Diese Disziplin zeigt sich als Methodologie, die sowohl in der theoretischen Linguistik als auch in vielen Anwendungsgebieten der Sprachwissenschaft, z.B. in der maschinellen Sprachverarbeitung und Übersetzung, in der Lexikographie (bei der Erstellung von Wörterbüchern), im Sprachunterricht, in der Erforschung von Sprachstörungen usw., eingesetzt werden kann (vgl. LEMNITZER, ZINSMEISTER, 2015, 14f.). Die Korpuslinguistik ist durch das Verwenden von authentischen Sprachdaten charakterisiert, die Erkenntnisse der Korpuslinguistik basieren auf natürlichen Äußerungen einer Sprache. Für diese wissenschaftliche Disziplin ist also wichtig, wie die Sprache tatsächlich verwendet wird.

Der Ausdruck **Textkorpus** bezeichnet generell eine Sammlung von schriftlichen Texten oder eine Sammlung von mündlichen Äußerungen, die schriftlich aufgezeichnet wurden. Die Daten des Textkorpus sind in der heutigen Zeit meistens digitalisiert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den *Primärdaten* (also aus den Texten selbst) und in der Regel auch aus den *Metadaten* (also aus den Zusatzinformationen zu einem Text als Ganzem wie z.B. Datum der Veröffentlichung, Verlag, Autor) und aus den *Annotationen* (also aus den Zusatzinformationen zu Elementen unterhalb der Textebene, also zu Textstellen oder –Passagen) (vgl. PERKUHN, 2012, 45). Typisch für in der Korpuslinguistik verwendete Korpora ist, dass sie nur zusammenhängende und vollständige Texte aus natürlichen Produktionssituationen beinhalten. Die Korpora unterscheidet voneinander die Gliederung nach dem *Medium* (geschriebene vs. gesprochene Sprache) und nach den *Sprachdomänen* (Sprachen oder Sprachausschnitten). Der andere Unterschied ist auch das, ob oder wie sie annotiert sind und ob es sich um universell verwendbare oder Spezialkorpora handelt. In der Sprachwissenschaft verwendete Korpora unterscheiden sich typischerweise aufgrund ihres Verwendungszwecks von Korpora und elektronischen Textsammlungen, die in anderen Disziplinen (z.B. in der Literaturwissenschaft oder in der Rechtswissenschaft) verwendet werden, in der Art der Metadaten und Annotationen, die sie enthalten (vgl. PERKUHN, ebd.) .

Ein Korpus soll eine Sprache (oder einen Dialekt oder Soziolekt etc.) **repräsentativ** abbilden. Die meisten Textkorpora setzen sich zum Ziel, möglichst universell bei der Erforschung einer gegebenen Sprache verwendbar zu sein. Im Idealfall sollten solche Korpora hinsichtlich möglichst vieler relevanter linguistischer Fragestellungen, Schlussfolgerungen und Verallgemeinerungen zulassen (vgl. PERKUHN, 2012, 47). Die Zusammenstellung von großen Korpora soll **ausgewogen** (engl. balanced) sein, also in einem bestimmten Verhältnis aus unterschiedlichen Textsorten bestehen. Für das DWDS-Kernkorpus (Digitales Wörterbuch der deutschen Sprache) war entlang der Dimension Textsorte eine Verteilung von 21,05% Gebrauchsliteratur, 28,42% Belletristik, 23,15% Wissenschaft und 27,36% Zeitungstexte geplant (vgl. PERKUHN, 2012, 48).

Die Sprachkorpora können als Hilfsmittel oder als Ausgangspunkt für verschiedene sprachwissenschaftliche Aufgaben dienen. Als Beispiel werden folgende linguistische Disziplinen ausgewählt:[1]

- Theoretische Linguistik: von der Überprüfung von Hypothesen bis zur automatischen Ermittlung grammatischer Regularitäten
- Lexikographie: Ermittlung von Worthäufigkeiten, Wendungen und typischen Verwendungskontexten, Sammlung authentischer Belege
- Grammatikographie: Belege für grammatische Strukturen, deren Häufigkeit und Verteilung
- Fremdsprachenunterricht: Analyse von Lernerfehlern, Ermittlung gebrauchshäufiger Phänomene, authentische Belege für Sprachverwendung
- Übersetzung: Überprüfung von Übersetzungsstrategien in Parallelkorpora.
- Computerlinguistik: automatische Übersetzung, Spracherkennung, etc.

2.2 Typen von Korpora

Die Korpora lassen sich nach mehreren formalen und inhaltlichen Kriterien kategorisieren. In diesem Kapitel werden die wichtigsten erwähnt.[2]

Man unterscheidet zum Beispiel zwischen **Papierkorpora** und **elektronischen Korpora**. Papierkorpora spielten vor allem in der Vergangenheit in der Wörterbuchschreibung eine wichtige Rolle. Anhand der Sammlungen wurden die Bedeutungen einzelner Wörter ausgemacht oder belegt. Die Papierkorpora waren manuell erstellt und ihre Herstellung war sehr aufwändig. Im Unterschied dazu stehen

[1] Die Aufzählung wurde von der Präsentation von Stefan Engelberg (Linguistische Methodenlehre, FS 2009, Uni Mannheim) übernommen. Die Präsentation ist abrufbar unter : http://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_LingMeth/Skript_05.pdf [abgerufen am 5.1.2017].

[2] Die Kategorisierung in diesem Kapitel wurde von Scherer (2006) übernommen , die auch unter https://de.wikipedia.org/wiki/Textkorpus abrufbar ist[abgerufen am 5.1.2017].

heutzutage die elektronischen Korpora zur Verfügung. Für die Nutzung der heute üblichen maschinenlesbaren Korpora braucht man entweder eine spezielle Software, wie beispielsweise COSMAS II[3], oder man kann etliche Korpora online verwenden, ohne solche Software auf dem eigenen Computer zu installieren.

Eine andere Gliederung stellen auch die **Teilkorpora** und **Referenzkorpora** dar. Teilkorpora sind solche Korpora, die nur einen Ausschnitt der Sprache dokumentieren. Es handelt sich zum Beispiel um solche Textkorpora, die nur Texte aus Tageszeitungen oder aus der Umgangssprache zu Verfügung stellen. Ein Referenzkorpus ist dagegen ein Textkorpus, das eine konkrete Sprache (also das Deutsche, das Englische, das Tschechische usw.) in ihrer Gesamtheit dokumentieren soll. Anhand eines Referenzkorpus können gültige Aussagen über das konkrete Sprachsystem generell formuliert werden. Referenzkorpora zielen nicht auf einen Ausschnitt der Sprache, aber es lassen sich in den Referenzkorpora verschiedene Textsorten finden (z.B. die belletristischen, wissenschaftlichen und populärwissenschaftlichen Texte, eine große Zahl von Zeitungstexten, diverse weitere Textsorten).

Man kann weitere Aspekte unterscheiden – **Statische Korpora** und **Monitorkorpora**. Statische Korpora sind abgeschlossen und werden nicht mehr erweitert. Es handelt sich z.B. um die Textkorpora mit den Werken eines verstorbenen Schriftstellers, um ein Korpus, das aus der Gesamtheit aller in einer ausgestorbenen Sprache vorhandenen schriftlichen Quellen besteht oder im Allgemeinen um ein Korpus, dessen Zusammensetzung sich mit der Zeit nicht verändert. Ein solches Korpus muss nicht nur Sprachbelege der alten Sprachen enthalten, es kann sich auch um ein Korpus der heutigen Sprache handeln, wichtig ist aber der Aspekt, dass die Daten im statischen Korpus abgeschlossen sind. Alte Sprachen, die man nur in wenigen Dokumenten oder gar nur fragmentarisch belegt finden kann, werden als „Korpussprachen“ bezeichnet, weil sie nur anhand dieses begrenzten Textkorpus rekonstruierbar sind und man kann sie nur anhand dieses Korpus beschreiben. Monitorkorpora hingegen sind solche Textkorpora, die immer erweitert werden und

[3] Die Software COSMAS II ist heutzutage online anwendbar, früher musste auch auf dem Computer installiert werden.

deren Datensammlung nicht abgeschlossen ist. Es kann sich um verschiedene Textsorten handeln.

Andere Typen von Korpora stellen die **Rohkorpora** und die **annotierten Korpora** dar. Unter Rohkorpora versteht man solche Textkorpora, die nur aus den rohen Sprachdaten bestehen. Die meisten Korpora bestehen aber auch aus sog. *Metadaten* und *Annotationen*. Die Metadaten enthalten die Zusatzinformationen zu einem Text als Ganzem wie z.B. Datum der Veröffentlichung, Verlag, Autor, Autorenschaft, Angaben über die Korpuserstellung. Den Begriff Annotation verwendet man – obwohl es sich bei Annotation streng genommen auch um Metadaten handelt – für Zusatzinformationen zu Elementen unterhalb der Textebene, also zu Textstellen oder –passagen. Im Bereich linguistischer Korpora wird der Begriff Annotation oft auch im Sinne von linguistischen Annotationen (d. h. für linguistische Kategorien, wie z.B. die Wortart) verwendet (vgl. PERKUHN, 2012, 45). Die Annotationen können in jedem Korpus unterschiedlicher Art sein. Gewöhnlich sind z.B. solche Korpora, in denen für jedes einzelne Wort zusätzlich die jeweilige Wortart angegeben wird, was für die linguistischen Zwecke sehr nützlich sein kann.

Weiter lassen sich auch solche Korpora finden, die Glossen enthalten, oder Korpora, die mit Angaben die Syntax der einzelnen Sätze betreffend versehen sind – solche Korpora werden auch als **Baumbanken** bezeichnet (analog zum Ausdruck „Datenbank"), da in diesen Korpora sogenannte syntaktische Baumstrukturen annotiert sind. Textkorpora der gesprochenen Sprache werden oft auch mit phonologischen Daten angereichert. Annotierte Korpora bieten verbesserte Forschungsmöglichkeiten im Gebiet der Linguistik, man kann dank der Annotation tiefschürfende und schnellere Analyse durchführen und präzisere Ergebnisse zu der gesuchten Spracherscheinung gewinnen.

Eine andere Gliederung betrifft die Anzahl der Sprachen, die in dem Korpus beinhaltet sind. Es gibt **einsprachige Korpora**, die Texte aus der Einzelsprache enthalten und **mehrsprachige Korpora**, die über Aussagen in zwei oder mehreren Sprachen verfügen. Entweder sind die Texte in der zweiten Sprache eine Übersetzung der Texte der ersten Sprache – solche Texte werden als **Parallelkorpora** bezeichnet –

oder die Sammlung der Texte der zweiten Sprache besteht im selben Ausmaß aus denselben Textsorten wie das Korpus der ersten Sprache (z. B. Zeitungsartikel zu denselben Themen). Sie werden **Vergleichskorpora** genannt. Dank der mehrsprachigen Korpora erleichtert sich vor allem die maschinelle Übersetzung, diese Korpora spielen auch für die Sprachlehrforschung eine Rolle. Die mehrsprachigen Korpora können auch ein wichtiger Beitrag für die automatische Erstellung eines zweisprachigen Wörterbuches sein.

2.3 Größe der Korpora

Eine entscheidende Rolle spielt auch die **Größe** eines Korpus. Ob die größeren oder kleineren Korpora für den Nutzer geeigneter sind, hängt davon ab, welche Fragestellungen und welche Ziele der Nutzer mit der Durchführung konkreter Analyse hat. In einigen Fällen können auch kleine Korpora nützliche Ergebnisse bringen (z.B. wenn man Belege für die Verwendung eines Wortes sucht), in anderen Fällen (wenn man z.B. ein Wort in Kombination mit einem semantischen Faktor untersuchen möchte) droht die Gefahr, dass man keine genügenden Belege findet – in solchen Fällen wären deutlich größere Korpora notwendig. Generell gilt aber immer ein geprägter Slogan von Robert Mercer „more data is better data“ (vgl. PERKUHN, 2012, 50). Vor allem in der Computerlinguistik werden möglichst große Daten ausgewertet, um allgemeine Aussagen über eine Sprache treffen zu können.

2.4 Die bekanntesten Korpora

Zu den bekanntesten sprachwissenschaftlichen Korpora zählen vor allem das **Brown Corpus** und das **British National Corpus** (BNC). Für Vorreiter der deutschen Korpuslinguistik kann man das Institut für Kommunikationswissenschaft und Phonetik (IKP) an der Universität Bonn und das Institut für Deutsche Sprache (IDS) in Mannheim halten. Heute sind als bedeutendste deutschsprachige Korpora besonders folgende zu nennen:

- das **Deutsche Referenzkorpus** (DeReKo) am Institut für Deutsche Sprache in Mannheim
- das Kernkorpus des **Digitalen Wörterbuchs der Deutschen Sprache** (DWDS) an der Berlin-Brandenburgischen Akademie der Wissenschaften, im Rahmen des Forschungsprojekts „DWDS“ wurde das größte ausgewogene Textkorpus der deutschen Sprache des 20. Jahrhunderts bereitgestellt
- das Korpus des Projekts **Deutscher Wortschatz** an der Universität Leipzig (vorwiegend Texte aus Online-Medien)
- das **Schweizer Textkorpus** an der Universität Basel

Es stellt sich auch die Frage, ob auch das **World Wide Web** als Korpus angesehen werden kann. Diese Frage ist aber unter Linguisten aus mehreren Gründen umstritten. Das erste Problem stellt die Dauerhaftigkeit dar – die Informationen im Internet ändern sich ständig und es ist unmöglich, die Anzahl von Wörtern zu ermitteln. Im Internet lassen sich unverlässliche Zitate finden und man kann oft schwierig oder gar nicht die Annotation ermitteln. Andere Nachteile stellen kleiner Kontext, zu viele Graphik, Absenz der Diakritik, unvollständige Sätze und Orthographie- oder Grammatikfehler dar. Die diachronische Analyse im Internet ist auch nicht möglich. Die Zusammensetzung der Daten ist nicht kontrollierbar, die Suchergebnisse sind immer unterschiedlich, man kann auch keine Kookkurrenzanalyse durchführen. Als Vorteil gelten die schnelle Suche und die große Datenmenge. Die Sprache im Internet muss

aber immer für Mittel zum Erwerb von Informationen und nicht für das Ziel gehalten werden.

Als das bekannteste deutsche Webkorpus kann das **deutsche Web as Corpus** (deWaC) betrachtet werden. DeWaC existiert seit 2006 und enthält mehr als 1,7 Milliarden Tokens. Nach deutschem Recht ist aber die Verwendung dieser Korpora nicht ganz unproblematisch, da sie ohne explizite Erlaubnis der Urheber aufgebaut wurden (vgl. PERKUHN, 2012, 46).

2.5 Corpus-based vs. corpus-driven

Im Rahmen der Korpuslinguistik müssen auch zwei Suchmöglichkeiten erwähnt werden – **corpus-based** und **corpus-driven**.

Vom corpus-driven bzw. korpusgeleiteten Paradigma spricht man im Fall, wenn die Hypothesen ausschließlich aus dem Textmaterial heraus generiert werden sollen. Diese Suchmöglichkeit kann z.B. helfen, Kookkurrenzen incl. Kollokationen und Phraseme zu entdecken. Die Textdaten dienen also nicht dazu, erst im Nachhinein Thesen oder Theorien zu bestätigen oder zu widerlegen. Die Textdaten stellen eigentlich den Ausgangspunkt dar, von dem aus Thesen abgeleitet und Theorien aufgestellt werden (vgl. PERKUHN, 2012, 20).

Dagegen sagt die Definition des Korpus-Begriffs, dass Korpora sowohl zur Ermittlung und Beschreibung sprachlicher Regularitäten als auch zur Überprüfung von vorher aufgestellten Hypothesen und Theorien dienen. In diesem Fall spricht man vom so genannten corpus-based bzw. korpusbasierten Paradigma. Diese Suchmöglichkeit überprüft die Existenz und Frequenzangaben der gesuchten Suchanfragen und je nach der Suchanfrage kann auch die Variabilität entdeckt werden. Die Synonyme und Antonyme können in der corpus-based Analyse nur zufällig entdeckt werden.

2.6 DeReKo und COSMAS II

Die Korpora geschriebener Gegenwartssprache des IDS bilden mit über 31,68 Milliarden Wörtern (Stand 08.03.2017) die weltweit größte linguistische Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit. [4]

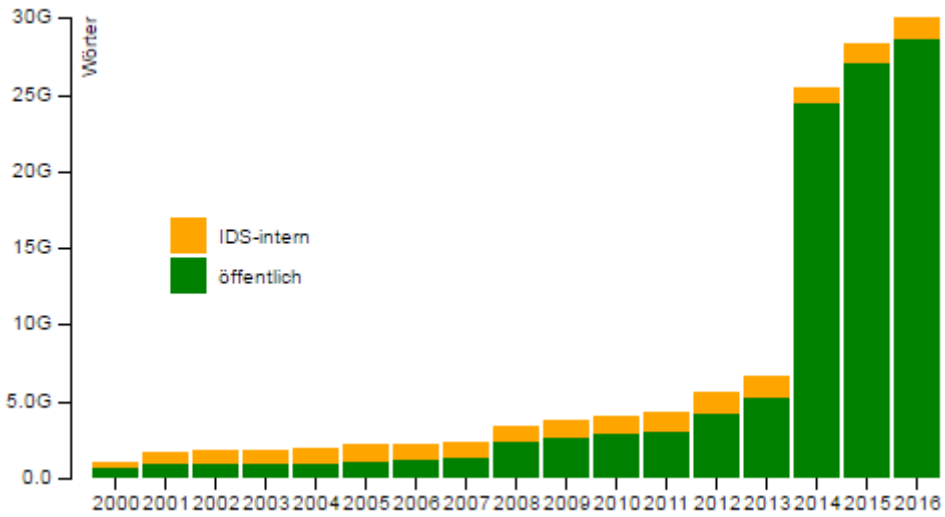


Abb. 1 – Die Größe der IDS-Korpora[5]

Die Größe der IDS-Korpora erlaubt[6]:

- seltene Fälle eines Wortgebrauchs ausfindig zu machen,
- die Vielfalt der deutschen Wortkompositionen einzubeziehen,
- mittels Statistiken (**Kookkurrenzanalyse**) starke Wortverbindungen, geläufige Assoziationen und syntaktische Muster des Gebrauchs von Wörtern zu identifizieren,

[4] Die Informationen in diesem Kapitel wurden aus http://www.ids-mannheim.de/cosmas2/projekt/einsteiger/was.html übernommen [abgerufen am 5.4.2017].
[5] Die Grafik wurde aus http://www1.ids-mannheim.de/kl/projekte/korpora/archiv.html übernommen [abgerufen am 5.4.2017].
[6] Die Aufzählung wurde aus http://www.ids-mannheim.de/cosmas2/projekt/einsteiger/wozu.html übernommen [abgerufen am 5.4.2017].

- zeitliche und kontextuelle Schwankungen in der Verwendung von Wörtern zu entdecken,
- neue Wörter oder neue Bedeutungen von vorhandenen Wörtern (Neologismen) zu extrahieren,
- Recherchen auf spezielle Textuntermengen einzugrenzen (Texte nach dem 11. Sept. 2001; Belletristik der 90er; etc.),
- das erste Auftreten von festen Wortverbindungen (z.B. *roter Faden*) oder Termini (z.B. *Bundeskanzlerin*) zu lokalisieren und deren Gebrauch zeitlich zu verfolgen,
- eine Erklärung zu erhalten für Wörter, die man in einem Wörterbuch vielleicht nicht findet,
- gesellschaftliche Trends anhand des Vokabulars zu erfassen,
- die Variabilität grammatikalischer Satzgefüge aufzufächern.

Die Sammlung wird immer kontinuierlich weiterentwickelt und enthält ausschließlich urheberrechtlich abgesichertes Material. Die Ergebnisse der Arbeit an diesen Korpora werden in regelmäßigen Abständen veröffentlicht, indem sie an das Projekt **COSMAS II** übergeben werden. COSMAS II steht für *Corpus Search, Management and Analysis System* und ist das Nachfolgesystem von COSMAS I. Über die Webapplikation COSMAS II ist **DeReKo** (Deutsches Referenzkorpus) öffentlich und kostenlos für die Nutzer zugänglich. COSMAS II kann nicht als eine Suchmaschine (wie z.B. Google o.ä.) verstanden werden, die die Inhalte von Servern im Internet durchsuchen kann. COSMAS II wird auch nicht zu kommerziellen Zwecken entwickelt und darf zu diesen Zwecken auch nicht genutzt werden.

Das immer wachsende Deutsche Referenzkorpus, das COSMAS II zugänglich macht, enthält belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Menge von Zeitungstexten und Texte vieler anderer Textarten aus Deutschland, Österreich und der Schweiz von 1772 bis heute.

Die COSMAS II - Korpora sind in **Archiven** organisiert. Archive sind eine Art Sammelstellen, denen Korpora zur Lagerung zugeführt werden. Es gibt insgesamt 18 Archive der COSMAS II - Korpora. Das Hauptarchiv ist **das Archiv W der**

geschrieben Korpora. Dieses Archiv ist am größten und umfasst Texte vom 18. Jahrhundert bis heute[7]. Dieses Archiv wird auch für die Zwecke der vorliegenden Diplomarbeit benutzt.

2.6.1 Beschreibung der Analyse in COSMAS II

Eine Recherche in COSMAS II kann wie folgt verlaufen[8]:

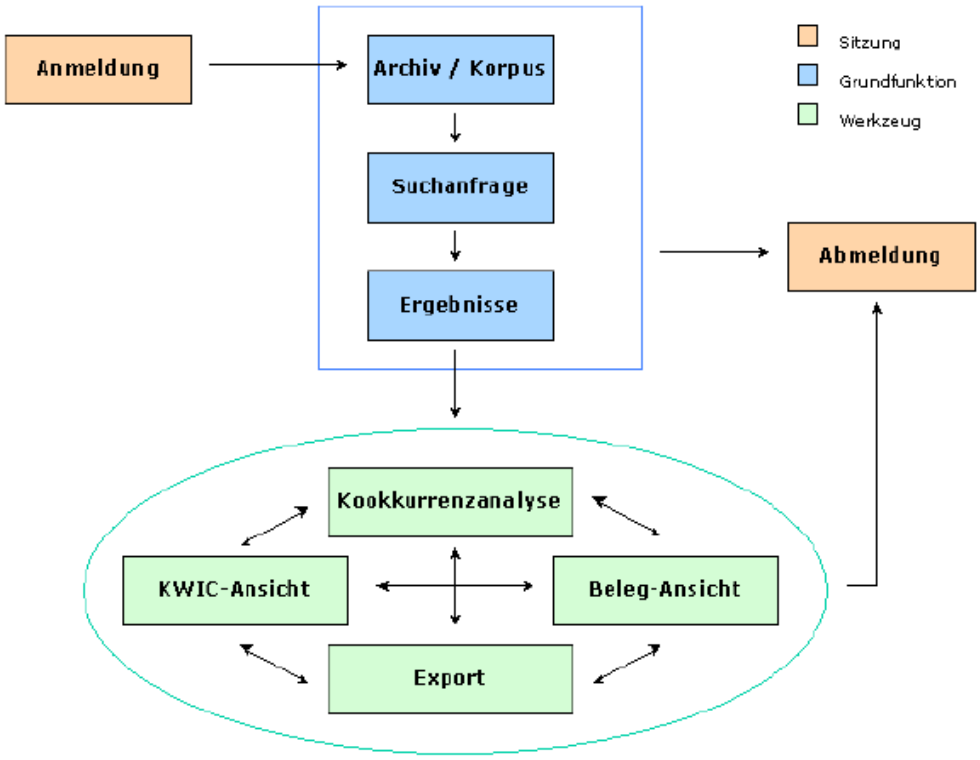


Abb. 2 – Struktogramm – COSMAS II

Dieses Struktogramm zeigt vereinfacht und in groben Zügen die vielfältigen Funktionalitäten von COSMAS II. Die Arbeit in COSMAS II lässt sich in zwei Arbeitsschritte - die **Grundfunktionen** (blau markiert) und die **Werkzeuge** (grün markiert) - zusammenfassen. Mit den Pfeilen wird auf dem Struktogramm ein Vor- oder

[7] Die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/projekt/referenz/archive.html übernommen [abgerufen am 27.2.2017].
[8] Das Struktogramm und die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/projekt/hilfe/quick.html übernommen [abgerufen am 27.2.2017].

Zurückspringen zwischen den verschiedenen Funktionalitäten im Rahmen einer Korpusanalyse symbolisiert.

Wie es auf der Graphik zu sehen ist, muss man nach der **Anmeldung** zuerst ein **Archiv** wählen (z.B. W – Archiv der geschriebenen Korpora) und innerhalb dieses Archivs wird ein sogenanntes virtuelles **Korpus** geladen. Als virtuelles Korpus kann entweder ein vordefiniertes Korpus gewählt werden (z.B. W-öffentlich – alle öffentlichen Korpora des Archivs W) oder ein neues Korpus definiert werden. Während einer Sitzung kann sowohl das Archiv als auch das Korpus gewechselt werden. Der nächste Schritt ist die Formulierung einer **Suchanfrage**.

Suchanfragen in COSMAS II werden u.a. gebildet aus[9]:

- Wörtern, Teilwörtern, Wortgrundformen,
- Wortklassen (z.B. Verb, Artikel) und grammatikalischen Mustern,
- Angaben zu Wort- und Satzabstand,
- Angaben zu Textbereichen (z.B. Überschriften) und Position (z.B. erstes Wort eines Satzes).

Mit der passenden Formulierung der Suchanfrage kann dem Nutzer eine Sammlung von Suchanfragebeispielen helfen, die im COSMAS II zur Verfügung steht. Je präziser diese Anfrage ist, desto konkretere **Ergebnisse** bietet das System an.

Ergebnisse in COSMAS II können u.a.[10]:

- nach Entstehungszeit, Erscheinungsland und Thematik sortiert werden,
- durch Statistiken auf häufig verwendete Gebrauchsmuster hin analysiert werden.

[9] Die Aufzählung wurde aus http://www.ids-mannheim.de/cosmas2/projekt/service/flyer10.pdf übernommen [abgerufen am 5.4.2017].
[10] Die Aufzählung wurde aus http://www.ids-mannheim.de/cosmas2/projekt/service/flyer10.pdf übernommen [abgerufen am 5.4.2017].

Falls für den Nutzer nur die Ergebnisse dieser einfachen Analyse nützlich sind, kann er schon in diesem Punkt die Analyse beenden. Sonst bietet die Analyse noch andere Werkzeuge an.

Die **KWIC-Ansicht** (Key Word In Context) ist eine zeilenweise Auflistung von Treffern in ihrem Kontext, in der die Suchbegriffe zentriert ausgerichtet und farblich hervorgehoben sind. In einer KWIC-Ansicht werden die Treffer und ihr jeweiliger Kontext in einer kompakten tabellenartigen Ansicht präsentiert.[11] Im Unterschied zur KWIC-Ansicht werden bei **Beleg-Ansicht** die Treffer nicht in einem auf eine Zeile begrenzten Ausschnitt angezeigt, sondern in einem breiteren Kontext, der bis drei Absätze groß sein kann. In diesem Fall spricht man auch von einer **Volltext-Ansicht**.

Wenn man die Suchanfrage erfolgreich durchführt und die Ergebnisse (oder andere Aspekte der Recherche) in eine Datei sichern will, stehen im System verschiedene Einstellungen zur Verfügung, mit denen man den **Export** individuell konfigurieren kann. So kann man vor der Abmeldung die Ergebnisse der durchgeführten Analyse speichern.

2.6.2 Kookkurrenzanalyse

Im Rahmen von COSMAS II steht den Nutzern auch eine korpusanalytische Methode zur Strukturierung von Belegmengen zur Verfügung – man spricht über sogenannte **Kookkurrenzanalyse** (KA).

Die **KA**[12]:

- ermöglicht das Aufdecken von signifikanten Regelmäßigkeiten bei der Verwendung von Wortkombinationen in den Korpora

[11] Die Definition wurde aus http://www.ids-mannheim.de/cosmas2/web-app/hilfe/seite/kwic/ übernommen [abgerufen am 5.1.2017].
[12] Die Aufzählung wurde aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 5.4.2017].

- wertet mit Hilfe mathematisch-statistischer Analyse- und Clusteringverfahren den definierbaren Kontext eines vorgegebenen Suchobjekts in beliebigen virtuellen Korpora aus

- liefert Hinweise auf systematisches gemeinsames Auftreten von Wörtern (Kookkurrenzen) und ein Maß für deren Affinität (Kohäsion)

- fasst Belege, die ähnliches Kookkurrenzverhalten des Wortes dokumentieren, zu Gruppen/Clustern zusammen

- strukturiert diese Belegmengen ggf. hierarchisch

- bietet eine entsprechende synoptische Präsentation der Belege

- erfasst neben binären Wortrelationen auch usuelle phrasale Muster bis hin zu (idiomatischen) Mehrworteinheiten.

Der Begriff Kookkurrenz soll so verstanden werden, dass er das rekurrente gemeinsame Vorkommen sprachlicher Einheiten in den Datenbeständen bewertend festhält (vgl. PERKUHN, 2012, 113). „Als Kookkurrenz wird das gemeinsame Vorkommen zweier oder mehrerer Wörter in einem Kontext von fest definierter Größe bezeichnet. Dabei sind Kookkurrenzen dort linguistisch interessant, wo das gemeinsame Auftreten der Wörter häufiger zu beobachten ist, als bei einer Zufallsverteilung aller Wörter zu erwarten wäre“ (KUNZE, LEMNITZER, 2007, 391f). Die **KA** kann also als ein Verfahren beschrieben werden, das bewertet, ob die Häufigkeit der sprachlichen Einheit in einer anderen Menge dem entspricht, was man erwarten darf, oder ob man nicht doch überzufällig viele Ereignisse beobachtet, die für eine Assoziation der Einheit mit einer Eigenschaft der Menge sprechen (vgl. PERKUHN, ebd.). Es lässt sich an einem Beispiel demonstrieren: In einem Korpus tritt Wort X tausendmal auf, Wort Y hundertmal, Wort Z zehnmal. Daraus folgt, dass die Kombination XY ist zehnmal so wahrscheinlich als die Kombination XZ. XY sollte zehnmal so oft vorkommen als XZ. Die Ergebnisse der KA können aber zeigen, dass XZ und XY gleich oft vorkommen.

Daraus lässt sich schlussfolgern, dass es eine besonders enge sprachliche Verbindung zwischen X und Z (über das Erwarten hinaus) gibt.[13]

Im Rahmen lexikalischer Analysen werden die Kookkurrenzen z.B. zur Bestimmung der Valenz und der semantischen Relationen (Hyperonymie, Synonymie, Antonymie usw.) benötigt.

Die Kookkurrenzanalyse ist eine Bereicherung für Korpuslinguistische Untersuchungen. Dank der KA können die Kollokationen und andere (auch neue) feste Wendungen identifiziert werden. „Eine Kollokation ist ein aus meist zwei sprachlichen Zeichen bestehender Ausdruck, in dem die beiden sprachlichen Zeichen in arbiträrer und konventionalisierter Form verbunden sind (z.B. *blonde Haare, ein heikles Thema*). Innerhalb der Kollokation kann man die Basis als semantisch autonomes Element (Haare, Thema) und den Kollokator (blond, heikel) als semantisch abhängiges Element unterscheiden“ (ENGELBERG, LEMNITZER, 2001, 391f). Kollokationen können in gewisser Weise als linguistisch interpretierte Kookkurrenzen verstanden werden.

Das IDS stellt die KA seit 1995 integriert dem komplexen Online-System COSMAS II zur Verfügung. Die Kookkurrenzanalyse ist auf beliebige COSMAS-Suchobjekte anwendbar mit optionaler Lemmatisierung, variabler Kontextgröße, ggf. automatischer Fokussierung auf den Kontext mit dem stärksten Kohäsionswert, variabler Zuverlässigkeit (d.h. Signifikanz des ersten Kookkurrenzpartners), variabler Granularität (d.h. Signifikanz der Kookkurrenzpartner, die für die Ermittlung von Mehrworteinheiten berücksichtigt werden), variabler Zuordnung von Belegen bei Mehrworteinheiten und Berechnung von syntagmatischen Mustern zu jedem Kookkurrenzcluster.[14] Im Rahmen dieser Diplomarbeit soll auch festgestellt werden, inwieweit diese Parameter die KA beeinflussen können.

[13] Das Beispiel wurde von der Präsentation von Stefan Engelberg (Linguistische Methodenlehre, FS 2009, Uni Mannheim) übernommen.
Die Präsentation ist abrufbar unter :
http://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_LingMeth/Skript_05.pdf [abgerufen am 27.2.2017].

[14] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 27.2.2017] .

2.7 CCDB - die Kookkurrenzdatenbank

Die Kookkurrenzdatenbank CCDB ist eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Die aktuelle Version der CCDB ist eingeschränkt öffentlich zugänglich unter http://corpora.ids-mannheim.de/ccdb/. Diese Datenbank dient als empirische Grundlage für die Formulierung von neuen sprachgebrauchsbezogenen linguistischen Hypothesen, deren Modellierung und Implementierung. [15]

Ganz vereinfacht lässt sich sagen, dass die CCDB über eine Sammlung von Kookkurrenzanalyseergebnissen verfügt, die miteinander verglichen wurden. Cyril Belica, der heutzutage für den Programmbereich Korpuslinguistik des IDS verantwortlich ist, hat im Jahre 2001 auf der Basis von DeReKo[16] eine Datenbank angelegt, in der die Ergebnisse von Analysen zu mehr als 220.000 Grundformen mit über 770.000 verschiedenen Kookkurrenzpartnern abgelegt sind. Die CCDB enthält für jedes Wort die Ergebnisse von bis zu fünf verschiedenen KA-n in Form von Hierarchien von ähnlichen Verwendungen. Diese Analysen wurden mit unterschiedlicher Parametereinstellung durchgeführt. Es werden bis zu 100.000 Verwendungen pro Wort und Analyse gespeichert.[17]

Es folgt daraus, dass es sich um eine statische Sammlung von Kookkurrenzanalyseergebnissen handelt, die im Unterschied zu der Kookkurrenzanalyse des IDS nicht kontinuierlich erweitert wird.

Die CCDB erforscht die Eigenschaften von Kohäsionsrelationen für die Weiterentwicklung von Korpusanalysenmethoden. Dank der CCDB lassen sich schnell und einfach die Informationen zum Kookkurrenzverhalten einzelner Lexeme finden.

[15] Die Definiton der CCDB wurde aus http://www1.ids-mannheim.de/fileadmin/ids/Downloads/flyer-ccdb.pdf übernommen [abgerufen am 2.3..2017].
[16] Umfang des Korpus: 2,2 Md. Textwörter.
[17] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 2.3..2017].

Aus diesem Grund eignet sich diese Datenbank auch als Hilfsmittel bei der lexikografischen Arbeit. [18]

Um die Kookkurrenzdatenbank gut verstehen zu können, muss auch der Terminus Kookkurrenzprofil erklärt werden. Das Kookkurrenzprofil eines Wortes ist die Gesamtheit der signifikantesten Kookkurrenzen zu diesem Wort. Die Kookkurrenzprofile geben uns eine Auskunft darüber, wie Wörter im Sprachgebrauch verwendet werden. Das heißt, die Kookkurrenzprofile informieren uns, wie fest diese Wörter im Sprachgebrauch eingebettet sind, welche andere Wörter sich in unmittelbarer Nähe befinden und mit welchen Wörtern sie besonders häufig vorkommen. Eine einzelne Kookkurrenz erfasst dabei einen Aspekt oder eine Nuance der Verwendung. Eine Kookkurrenz entspricht einem Teil, das einen Beitrag zum Gesamtbild des Wortes leisten kann. Ein Kookkurrenzprofil erfasst dagegen viele wichtige Aspekte oder Nuancen der Verwendung eines Wortes (vgl. PERKUHN, 2012, 127). In der CCDB hat man die Möglichkeit, ein solches Kookkurrenzprofil des Wortes zu erstellen.

[18] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 2.3..2017].

3 Beschreibung der Kookkurrenzanalyse und der Analyse in der CCDB

In diesem Kapitel werden die Analysen (die KA und die Analyse in der CCDB) praktisch in Form der Abbildungen beschrieben. Es soll hier gezeigt werden, wie sich die Analysen durchführen lassen, wie sich die beiden Ergebnistabellen (KA und CCDB) auf den ersten Blick voneinander unterscheiden und wie man sie dechiffrieren soll. Es soll auch erwähnt werden, welche Einstellungen die beiden Analysen anbieten und inwieweit der Nutzer die Ergebnisse der beiden Analysen beeinflussen kann. Die Einstellungen, die dem Nutzer im Rahmen der Analyse zur Verfügung stehen, werden in diesem Kapitel detailliert beschrieben und untereinander verglichen.

3.1 Beschreibung der Kookkurrenzanalyse des IDS

Wie schon gesagt wurde, kann man sich in COSMAS II die Kookkurrenzen zu einem Suchausdruck berechnen lassen. Nach der Anmeldung bei COSMAS II und nach der Auswahl der Korpora kann man schon eine Suchanfrage ins System eingeben. Am Anfang wurde eine simple Abfrage nach einem bestimmten Wort formuliert, zu diesen Zwecken wurde das Verb *bestehen* (in der lemmatisierten Form *&bestehen*) ausgewählt. Nach der Bewilligung aller Wortformen (die kleinen und großen Buchstaben usw.), wird schon die Ergebnisliste angezeigt. Man wählt dann im WWW-Client oberhalb der Ergebnisliste den Reiter KA.

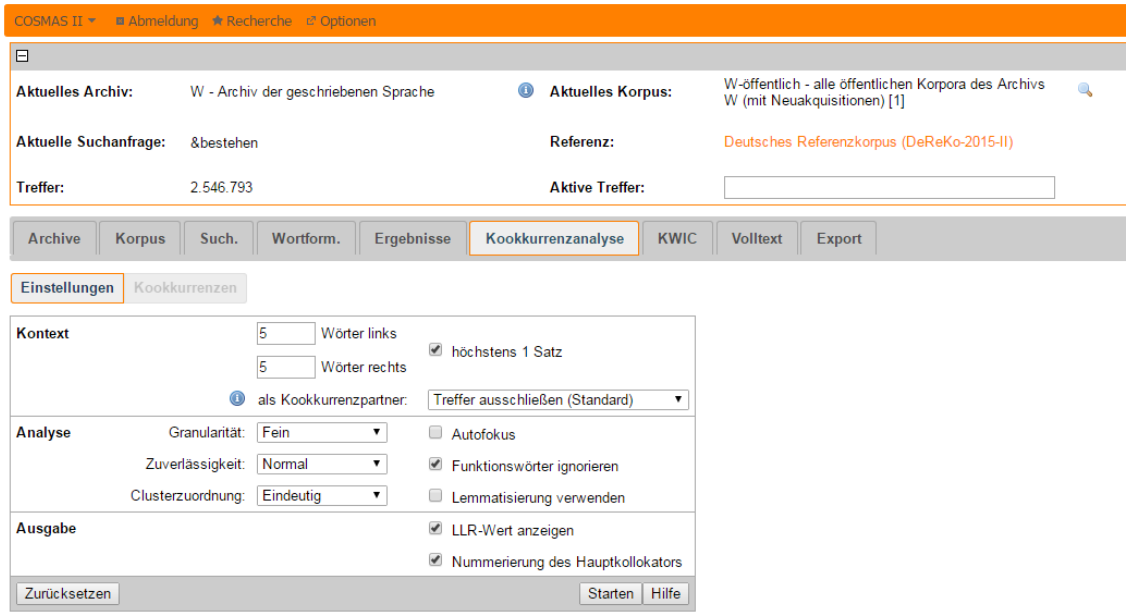


Abb. 3 – Wahl der KA

Im angezeigten Dialogfenster können genauere Einstellungen zur Analyse gemacht werden. Die wichtigsten Angaben betreffen den Bereich *Kontext*. Hier kann man entscheiden, wie viele Wörter nach links und nach rechts für die Analyse beachtet werden sollen. Also wie viele Wörter nach links und rechts überhaupt als mögliche Kookkurrenz-Partner zum Suchwort in Frage kommen. Das Ergebnis dieser Suche ist dann eine Liste von Wörtern aus der Umgebung des Bezugswortes – in diesem Fall des Verbs *bestehen*. Im Bereich *Kontext* findet man auch die Option *als Kookkurrenzpartner Treffer ausschließen, nur Suchwörter ausschließen* oder *Treffer samt Suchwörtern zulassen.*

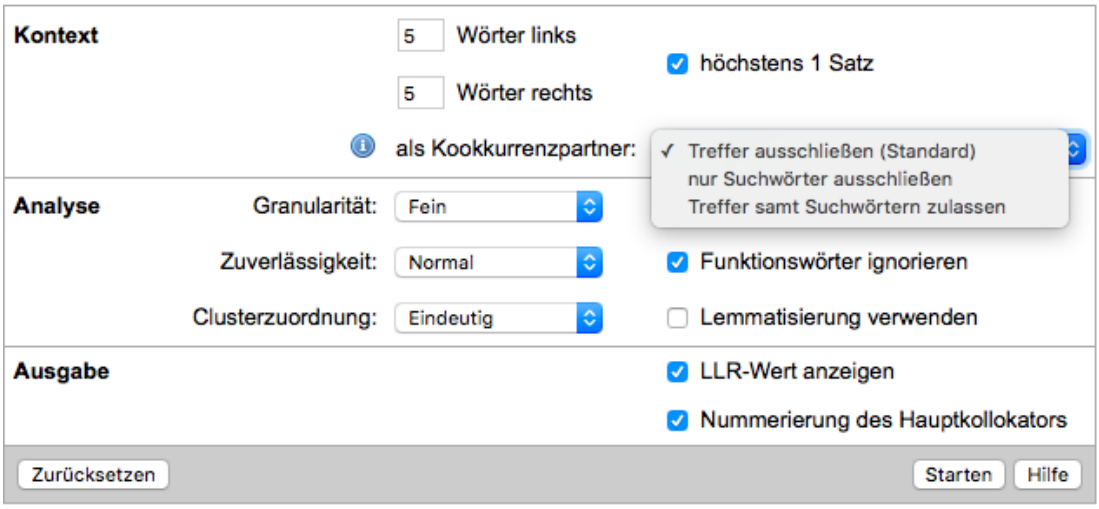


Abb. 4 – Einstellung der Parameter der KA

Der Optionswert *als Kookkurrenzpartner Treffer ausschließen* bedeutet, dass alle im eingestellten Kontext vorkommenden Treffer von der Analyse ausgeschlossen werden und als Kookkurrenzpartner nicht in Frage kommen. In dieser Standard-Einstellung werden alle im Kontext von 5 Wörtern rechts und 5 Wörtern links (je nach der Einstellung) gefundenen Treffer von der Analyse ausgeschlossen, so dass sie im Analyseergebnis auch nicht als Kookkurrenzpartner erscheinen. Das ist deshalb sinnvoll, weil die Analyse die Kookkurrenzpartner um die vorgegebene Wortsequenz herum ermitteln soll. Wenn man *nur Suchwörter als Kookkurrenzpartner ausschließt*, dann werden nur die in der Suchanfrage angegebenen Suchwörter und deren Expansionen von der Analyse ausgeschlossen. Das betrifft z.B. die Expansionen einer Grundform oder die Varianten eines Wortes bezüglich Groß-, Kleinschreibung und diakritischen Zeichen. Diese Expansionen treten also im Ergebnis der KA als Kookkurrenzpartner nicht auf. Wenn man die Option *als Kookkurrenzpartner Treffer samt Suchwörtern zulassen* auswählt, werden alle Wörter im eingestellten Kontext, bis auf das im KWIC als erstes hervorgehobene, in der Analyse berücksichtigt und sind als Kookkurrenzpartner zugelassen.[19]

In demselben Dialogfenster kann man auch die Einstellungen im Bereich der *Analyse* ändern. Für diesen Bereich sind vor allem die Begriffe *Granularität, Zuverlässigkeit* und *Clusterzuordnung* relevant.

Der Parameter *Zuverlässigkeit* ermöglicht dem Nutzer drei Stufen zu wählen – *hoch, normal, analytisch*. Mit *hoch* (nur starke Abweichungen sind relevant) maximiert man die Zuverlässigkeit, mit *analytisch* (schwache Abweichungen sind relevant) erreicht man die maximale Ausbeute. Es geht darum, ob beim Identifizieren von primären Partnerwörtern eher Wert auf höhere Präzision oder auf eine größere Vollständigkeit gelegt werden soll. Mit *hoch* bekommt man also eine kürzere Liste von Partnerwörtern, die aber zuverlässigere Kollokatoren sind. Mit *analytisch* erscheint eine lange Liste, an deren Ende die Kollokatoren stehen, die nicht so zuverlässig und signifikant sind.

[19] Die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/web-app/hilfe/seite/kookkurrenz/kook-opt-partner.html übernommen [abgerufen am 27.2.2017] .

Der Parameter *Granularität* gibt an, wie intensiv die Mehrwortgruppen gesucht werden sollen, d.h. wie viele der nach Signifikanz sortierten Kookkurrenzpartner als möglicher Kandidat eines Kookkurrenzpartners in Frage kommen. Es sieht praktisch so aus, dass der Nutzer auswählt, ob lieber sichere Aussagen über kurze Kombinationen oder auch etwas unsicherere Aussagen über eventuell komplexere Kombinationen gewünscht sind (vgl. PERKUHN, 2012, 113). Eine *sehr grobe* Granularität fokussiert die Analyse auf Schlagwörter, eine *feine* Granularität spürt Ausdrücke auf. Mit *sehr grob* bekommt man also eine Liste mit primären Kollokatoren, mit *fein* ist jeder primäre Kollokator noch weiter auf sekundäre (evtl. tertiäre) Kollokatoren segmentiert.

Die KA fasst Gruppen von Texten zusammen, die eine bestimmte Menge von Wörtern innerhalb eines Textfensters gemeinsam haben, diese Gruppen werden als (Text-)Cluster genannt. Je nach Einstellung des Parameters *Clusterzuordnung* umfasst ein Cluster alle (*mehrfach*) oder nur einen Teil (*eindeutig*) der verantwortlichen Textstellen (vgl. PERKUHN, 2012, 119). Bei *mehrfach* werden also Belege in alle relevanten Kollokationscluster eingefügt, bei *eindeutig* werden Ambiguitäten zugunsten des stärksten Kollokationsclusters aufgelöst. Mit *mehrfach* kann also ein Beleg bei mehreren Kollokatoren erscheinen, bei *eindeutig* ist der Beleg nur bei dem stärksten Kollokator zu sehen.

Im Bereich der Analyse lassen sich noch drei Einstellungen finden, und zwar *Autofokus, Funktionswörter ignorieren* und *Lemmatisierung verwenden*.

Ohne Autofokus wird der gesamte eingestellte Kontext betrachtet, mit Autofokus werden alle denkbaren zusammenhängenden Kontexte innerhalb des vorgegebenen Kontextes ausgewertet und es wird derjenige ausgewählt, der den höchsten Signifikanzwert aufweist.[20] Solche Analyse benötigt natürlich längere Zeit, aber der Nutzer gewinnt eine zusätzliche Information über die positionellen Präferenzen der primären Partnerwörter. Es muss bei kontrastiven Analysen darauf geachtet werden, dass diese zusätzliche Information eine neue, zu kontrollierende Variabilität in das Analyseergebnis hineinbringt (vgl. PERKUHN, 2012, 120).

[20] Die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/web-app/hilfe/seite/kookkurrenz/kook-opt-partner.html übernommen [abgerufen am 27.2.2017] .

Im Rahmen der Einstellung kann man auch die Funktionswörter aus der Analyse ausschließen (*Funktionswörter ignorieren*). Das System bekommt eigentlich den Befehl, ob auch Funktionswörter (z.B. Präpositionen, Artikel) und Satzzeichen als Kollokatoren gesucht werden sollen. Funktionswörter kommen sehr häufig in der Sprache vor und aus diesem Grund sind sie auch sehr stark in den Treffermengen vertreten. Diese Tatsache kann auch das Verdrängen anderer eventuellen Partner verursachen. Falls die Funktionswörter in die Analyse einbezogen werden, können syntaktische Phänomene interessantere Erkenntnisse überdecken. Für manche Wortverbindungen können aber gerade die Funktionswörter von Bedeutung sein (z.B. *bestehen in etw.*).[21]

Die Analyse erlaubt auch, die *Lemmatisierung* zu *verwenden*. Die Einstellung zur Lemmatisierung steuert, ob in der Liste einzelne Wortformen betrachtet werden oder ob die Einträge in der Liste, die als zu derselben Grundform gehörig betrachtet werden, gemeinsam als eine Einheit bewertet werden sollen (vgl. PERKUHN, 2012, 118). Dank diesem Parameter kann man entscheiden, ob als Kollokatoren Lemmata oder Textwörter gesucht werden sollen. Wenn man die Lemmatisierung einschaltet, dann werden die Wortformen nicht einzeln ausgewertet, sondern alle Wortformen, die auf dasselbe Lemma zurückgeführt werden können, werden zusammen ausgewertet. Dieser Schritt wird nur für die Wortformen in dem Kontext des Bezugswortes angewandt, nicht für das Bezugswort selbst. Ob dieses als Wortform oder Lemma gehandhabt werden soll, entscheidet man am Anfang bei der Formulierung der Suchanfrage.[22]

Die letzten zwei Einstellungen, die man in der KA auswählen kann, betreffen den Bereich der Ausgabe. Es handelt sich um die praktischen Funktionen *LLR-Wert anzeigen* und *Nummerierung des Hauptkollokators*. Mit der Option *LLR-Wert anzeigen*, wählt der Nutzer, ob der interne Wert für die ermittelte Stärke der lexikalischen Kohäsion angezeigt werden soll. Der LLR-Wert ist eine statistisch berechnete Größe der Affinität des Kookkurrenzpartners zum Bezugswort und die Kollokatoren in der KA sind nach dem LLR-Wert absteigend geordnet.

[21] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/misc/tutorial.html übernommen [abgerufen am 5.1.2017].
[22] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/misc/tutorial.html übernommen [abgerufen am 5.1.2017].

3.2 Anschauliche Durchführung der Kookkurrenzanalyse des IDS

In diesem Teil der Diplomarbeit wird beschrieben, welche Ergebnisse die KA dem Nutzer bringen kann und wie er diese Ergebnisse verstehen und dechiffrieren soll.

Der Ausgangspunkt für eine KA sind die vereinigten Kontexte aller Trefferobjekte einer Suchanfrage. Nach der Korpusauswahl muss man zunächst eine Suchanfrage formulieren. Bei der Formulierung muss darauf geachtet werden, dass die Wortform(en), die Gegenstand der Untersuchung sein soll(en), den eindeutigen Kern des Trefferobjektes ausmachen.[23]

Generell lässt sich sagen, dass die Anzahl der Treffer nicht zu klein sein darf, damit die statistischen Analysen verlässliche Aussagen liefern. Es gilt aber auch, dass wenn der Nutzer für eine Suchanfrage die Zufallsauswahl einschaltet oder die Treffermenge eine interne Obergrenze überschreitet, wird eine zufällig reduzierte Treffermenge als Ergebnis zurückgeliefert. Der Suchbegriff *bestehen* wurde in *W-öffentlich* gesucht und es wurden mehr als zwei Millionen Ergebnisse gefunden. Wenn die Anzahl der Kookkurrenzen für das System zu groß ist, wird die unreduzierte Trefferzahl nicht bestimmt, da die Ausführungszeit für diese Datenmenge zu viel Zeit beanspruchen würde. Die Reduzierung kann für den Nutzer auch günstig sein, denn ab einer Treffermenge von einigen Millionen Treffern werden die von der KA durchgeführten Berechnungen so aufwändig, dass die Durchführung der Analyse auch einige Stunden dauern kann.[24]

[23] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/misc/tutorial.html übernommen [abgerufen am 5.1.2017].
[24] Die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/server/themen/kook-anmerkungen.html übernommen [abgerufen am 5.1.2017].

KA - Grundeinstellung

Wenn man die KA durchführt, ohne die Einstellungen zu ändern, bietet das System automatisch folgende Parameter an:

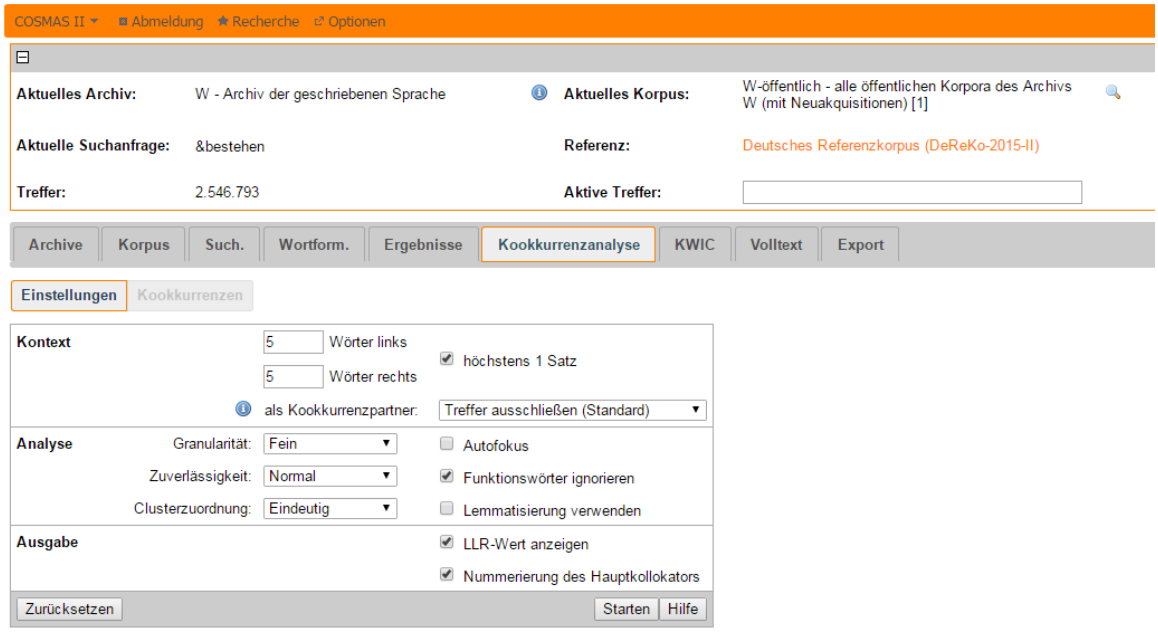


Abb. 5 – Grundeinstellung der KA – das Verb *bestehen*

Mit einem Klick auf „Starten“ fängt die Analyse an, die ungefähr 15 Minuten dauert. Als Ergebnis dieser Bewertung bekommt der Nutzer eine Liste von Wörtern aus der Umgebung des Bezugswortes (*bestehen*), mit ihren Häufigkeiten und ihrer statistischen Bewertung. Die auffälligen Kookkurrenzen werden als *primäre Partnerwörter* genannt (in diesem Fall *Gefahr*).

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	347478	1	1	**Gefahr** akute unmittelbare	100% besteht ... akute Gefahr ... unmittelbare
⊞			2	1	Gefahr akute Andernfalls	100% Andernfalls bestehe ... akute Gefahr
⊞			3	1	Gefahr akute Sonst	100% Sonst bestehe ... akute Gefahr
⊞			17	14	Gefahr akute Leib	64% Es bestehe [...] akute Gefahr für Leib und
⊞			19	2	Gefahr akute Grundwasser	50% bestand akute Gefahr ... Grundwasser
⊞			922	903	Gefahr akute	33% Es besteht [die|keine] akute Gefahr dass

Abb. 6 – Ergebnis der KA – *bestehen* – Kollokator *Gefahr*

Am Anfang jeder Konkordanz (Zeile der Ergebnisliste) ist ein Symbol „+“ zu sehen, wenn man dieses Symbol anklickt, bekommt er eine Möglichkeit, den Teil des Satzes im Kontext zu sehen, entweder in einer KWIC-Ansicht oder in einer Volltext-Ansicht, wo man auch die bibliographischen Angaben zum Text finden kann.

Gleich hinter der Nummerierung des Hauptkollokators erscheint das LLR. Die Liste ist nach dem LLR-Wert „Log Likelihood Ratio“ geordnet. Der LLR-Wert ist eine statistisch berechnete Größe der Affinität des Kookkurrenzpartners zum Bezugswort (hier 347478). Es handelt sich eigentlich um die Verbindungsstärke und es gilt, dass sich die häufigsten Kollokatoren in der Liste ganz oben finden lassen und die Reihung der Kollokatoren eine absteigende Tendenz hat.

Wie es auf der Abbildung 6 deutlich zu sehen ist, gibt es im Rahmen eines Clusters mehrere Partnerwörter. In der Spalte von Kookkurrenzen findet man die primären (hier *Gefahr*) und sekundären (hier z.B. *akute*) Partnerwörter. Mit dem Symbol „#“ werden die primären, in fetter Schrift wiedergegebenen, Partnerwörter nummeriert. In diesem Fall handelt es sich vor allem um die Substantive (*Gefahr, Möglichkeit, Handlungsbedarf, Zweifel, Verdacht* usw.), seltener um die Verben (*feiert, feierte* usw.) und um die Adjektive (*zehnjähriges, 50-jähriges, 25-jähriges, 100-jähriges* usw.).

In der Nebenspalte sieht man die typischen syntagmatischen Muster, in denen die Wörter zusammen auftreten (hier z.B. *besteht ... akute Gefahr ... unmittelbare*). Das syntagmatische Muster erfasst typische Realisierungen (lexikalische Umgebung) sowie syntaktische Struktur. Für die Angabe des syntagmatischen Musters werden keine tiefergehenden statistischen Auswertungen vorgenommen, es werden lediglich die Häufigkeiten innerhalb der Treffermenge ausgewertet. Das syntagmatische Muster beschreibt relative Reihenfolge der Clusterelemente (Bezugswort und Kookkurrenzpartner) - jeweils in einer bestimmten Ausprägung (nicht-lemmatisierte Wortformen) und die unbestimmte Füllung der Lücken bzw. bestimmte Füllung der Lücken mit Grad der Bestimmtheit.[25] Die Prozentangabe in dieser Spalte bezieht sich auf die Häufigkeit der Reihenfolge der konkreten Wortformen, die bei der Kookkurrenzanalyse beteiligt waren (vgl. PERKUHN, 2012, 122). Allgemein lässt sich sagen, dass diese Spalte typische Konstruktionen der Kookkurrenzen zeigt. Die Beispiele für syntagmatische Muster (Abb. 6) lassen sich so umschreiben: Bezogen auf die Anzahl der Textstellen wird in 100% der Fälle das Wort *Gefahr* eine Position nach *akute* beobachtet. Lücken zwischen den beteiligten Wörtern sind unbestimmt (als „…“) verzeichnet. In der vierten Konkordanz sieht man ein anderes Beispiel *Es bestehe*

[25] Die Definition des syntagmatischen Musters wurde aus http://www1.ids-mannheim.de/kl/misc/tutorial.html übernommen [abgerufen am 5.1.2017].

[…]akute Gefahr für Leib und, in solchem Fall sollen die Lücken als optional betrachtet werden. Die eckigen Klammern des Ausdrucks zwischen *bestehe* und *akute Gefahr* deuten an, dass die beiden Wörter nicht zwingend unmittelbar nacheinander folgen müssen und dass es sich auch optional ein Ausdruck dazwischen schieben kann.

Der nächste Kollokator für das Verb *bestehen* ist das Substantiv *Möglichkeit* (hier auf dem zweiten Platz mit LLR 288927). Bei diesem Kollokator lässt sich folgendes syntagmatisches Muster finden:

Möglichkeit Anschließend Stadtbummel	100% Anschließend besteht die Möglichkeit zu|zum einem Stadtbummel

Abb. 7 – Ergebnis der KA – *bestehen* – Kollokator *Möglichkeit*

In diesem Beispiel befindet sich unmittelbar vor *Stadtbummel* keine dominierende Besetzung. Neben der Form *zu* findet man auf derselben Position die Form *zu einem*. Die mögliche alternative Besetzung der Position erfasst beide Varianten, sowohl *Anschließend besteht die Möglichkeit zu Stadtbummel*, als auch *Anschließend besteht die Möglichkeit zu einem Stadtbummel*.

KA - Funktionswörter

Da in dieser KA des Wortes *bestehen* die Funktionswörter aus der Durchführung der Analyse ausgeschlossen waren, versuchte die Verfasserin die Analyse nochmals durchzuführen, ohne die Funktionswörter zu ignorieren. Am Anfang dieser Analyse wurde vor allem die Frage gestellt, ob mithilfe der KA auch die Valenz des Verbs bestimmt werden kann.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	863680	515115	515115	**aus**	40% besteht […] aus
⊞	2	567566	515120	5	**darin** Aufgabe Schwierigkeit	40% Schwierigkeit der Aufgabe bestand darin
⊞			515125	5	darin Aufgabe Herausforderung	60% Herausforderung der Aufgabe besteht […] darin
⊞			515126	1	darin Aufgabe Besonderheit	100% Besonderheit … Aufgabe bestand darin
⊞			515127	1	darin Aufgabe Vorteil	100% Aufgabe besteht darin … Vorteil
⊞			515137	10	darin Aufgabe Kunst	30% Aufgabe der p Theaterpädagogik besteht darin die Sprache der Kunst
⊞			521994	6857	darin Aufgabe	44% Die Aufgabe […] besteht […] darin die …
⊞			521995	1	darin Schwierigkeit Herausforderung	100% Schwierigkeit … Herausforderung … besteht darin
⊞			521996	1	darin Schwierigkeit Trick	100% Schwierigkeit besteht … darin … Trick
⊞			521997	1	darin Schwierigkeit Kunst	100% Schwierigkeit … Kunst bestehen darin
⊞			523459	1462	darin Schwierigkeit	58% Die Schwierigkeit […] besteht […] darin dass die
⊞			523460	1	darin Hauptaufgabe Kunst	100% Hauptaufgabe besteht darin … Kunst
⊞			524477	1017	darin Hauptaufgabe	41% Die Hauptaufgabe […] besteht […] darin die …
⊞			524481	4	darin Herausforderung Kunst	25% darin besteht … Herausforderung … Kunst
⊞			526727	2246	darin Herausforderung	54% Die Herausforderung […] besteht […] darin die …
⊞			526728	1	darin Besonderheit Kunst	100% Kunst besteht darin … Besonderheit
⊞			527378	650	darin Besonderheit	74% Eine|Die Besonderheit [des|der …] besteht […] darin dass …
⊞			527379	1	darin Vorteil Kunst	100% Kunst besteht darin … Vorteil
⊞			528911	1532	darin Vorteil	76% Der Vorteil […] besteht […] darin dass …
⊞			529295	384	darin Hauptproblem	58% Das Hauptproblem […] besteht […] darin dass die
⊞			529296	1	darin Trick Kunst	100% Trick … Kunst bestehen darin

Abb. 8 – Ergebnis der KA – *bestehen* – incl. Funktionswörter

Nach der Durchführung der KA ist schon auf den ersten Blick sichtbar, dass die Analyse ganz andere Ergebnisse als die vorherige Analyse geliefert hat. Während in der vorherigen Analyse die gewöhnlichsten Kollokatoren Substantive waren, in dieser Analyse sind die ersten zwei häufigsten Kollokatoren die Präposition *aus* (mit LLR 863680) und das Adverb *darin* (mit LLR 567566). Erst auf dem dritten Platz folgt das Substantiv *Gefahr* (mit LLR 347478), das in der vorherigen Analyse auf dem ersten Platz stand. Die Kollokatoren mit der stärksten Verbindungsstärke waren wegen der Einstellung aus der vorherigen Analyse ausgeschlossen.

Die durchgeführte Analyse hat bewiesen, dass sie u. a. zur Bestimmung der Valenz geeignet ist. Die KA hat die regierten Präpositionen bestimmt, die sich am häufigsten mit dem Verb *bestehen* verbinden, es handelt sich um die Präpositionen *aus, auf* und *in*, wobei die Präpositionen *auf* und *in* insbesondere als präpositionale Komplementsätze vorkommen. Im Unterschied dazu kommt die Präposition *in* oft in

Interrogativsätzen vor. Als Beleg dienen die folgenden syntagmatischen Muster, in denen die Wörter zusammen auftreten.

Abb. 9 – Ergebnis der KA – *bestehen* – Kollokator *aus*

Abb. 10 – Ergebnis der KA – *bestehen* – Kollokator *darin*

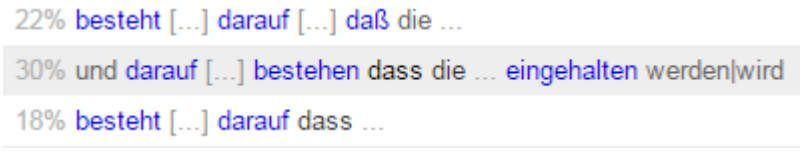


Abb. 11 – Ergebnis der KA – *bestehen* – Kollokator *darauf*

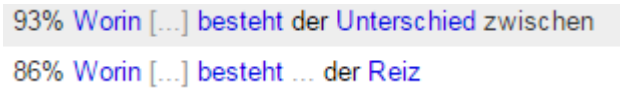


Abb. 12 – Ergebnis der KA – *bestehen* – Kollokator *Worin*

KA – Autofokus

Für die Durchführung der nächsten Analyse wurde dieselbe Suchanfrage eingegeben (*&bestehen*), nur die Einstellung für die KA wurde geändert. Diesmal sollte der Parameter *Autofokus* untersucht werden. Wenn man die Analyse mit Autofokus durchführt, ermittelt sie die typische Stellung der Kollokatoren im Kontext. Der Nachteil von dieser Parametereinstellung ist, dass die Analyse viel zeitaufwändiger ist. Die Funktionswörter wurden in dieser Analyse nicht ignoriert.

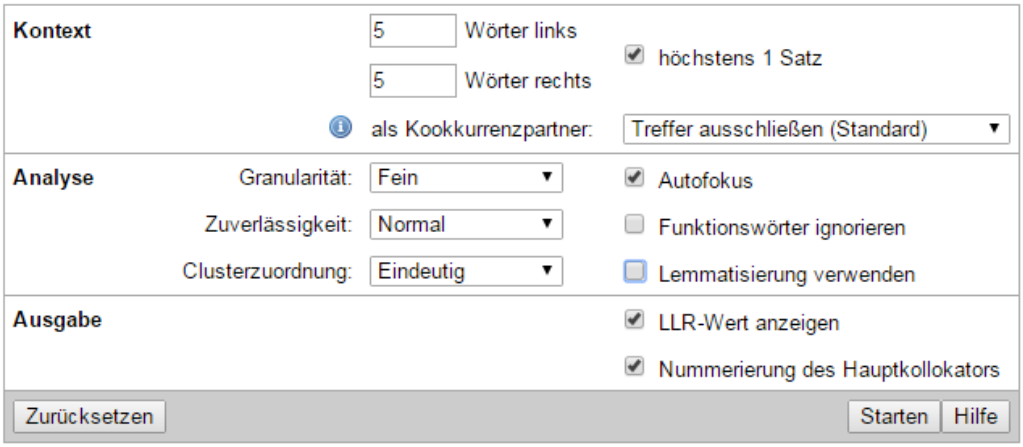


Abb. 13 – Einstellung der KA – mit *Autofokus*

Neben den Spalten, die in den vorherigen Analysen beschrieben waren, liefert die Analyse noch zwei Spalten, die gerade den Parameter *Autofokus* betreffen – *links* und *rechts.*

	#	LLR	kumul.	Häufig	links	rechts	Kookkurrenzen	syntagmatische Muster
⊞	1	1186526	515115	515115	1	1	**aus**	40% besteht [...] aus
⊞	2	652830	515120	5	-1	4	**darin** Aufgabe Schwierigkeit	40% Schwierigkeit der Aufgabe bestand darin
⊞			515125	5	-1	4	darin Aufgabe Herausforderung	60% Herausforderung der Aufgabe besteht [...] darin
⊞			515126	1	-1	4	darin Aufgabe Vorteil	100% Aufgabe besteht darin ... Vorteil
⊞			515136	10	-1	4	darin Aufgabe Kunst	30% Aufgabe der p Theaterpädagogik besteht darin die Sprache der Kunst
⊞			515137	1	-1	4	darin Aufgabe Besonderheit	100% Besonderheit ... Aufgabe bestand darin
⊞			515138	1	-1	4	darin Aufgabe Nachteil	100% Nachteil darin besteht ... Aufgabe
⊞			515139	1	-1	4	darin Aufgabe Kunststück	100% Kunststück ... Aufgabe ... darin bestehen

Abb. 14 – Ergebnis der KA – *bestehen* – Parameter *Autofokus*

Autofokus bestimmt, ob der gesamte eingestellte Kontext betrachtet wird oder nicht. Mit Autofokus werden alle zusammenhängenden Unterkontexte innerhalb des vorgegebenen Kontextes ausgewertet. Es wird dann der Kontext ausgewählt, der den höchsten Signifikanzwert aufweist.[26] Der Nutzer bekommt eine zusätzliche Information über die positionellen Präferenzen der primären Partnerwörter. Der Nutzer kann für jede Kookkurrenz den Autofokusbereich anzeigen, indem er den Zeiger auf der gewünschten Kookkurrenzzeile stillhält. Der erscheinende Text ist die verdeutschte Zusammenfassung der Felderwerte *links* und *rechts*.

Bei der Anzeige werden die Werte der linken und rechten Intervallgrenze zu einem Cluster angegeben. Die Notation unterscheidet mit vorangestelltem Minus- bzw. Pluszeichen zwischen linkerhand bzw. rechterhand des Bezugswortes (vgl. PERKUHN, 2012, 120). Es ist wichtig zu erwähnen, dass sich die Autofokus-Angabe auf das primäre Partnerwort bezieht. Diese Angabe sagt über die Anordnung der anderen Partnerwörter nichts aus.

Von dem ersten Beispiel (Abb. 15) lässt sich also ablesen, dass die typische Stellung im Kontext für *darin* das erste Wort links bis das vierte Wort rechts ist. Am zweiten Beispiel (Abb. 16) ist die Position fest angegeben, die typische Stellung im Kontext für *ihr* ist das zweite Wort links vor dem Bezugswort.

Abb. 15 – Ergebnis der KA – *bestehen* – Kollokator *darin* (Autofokus)

Abb. 16 – Ergebnis der KA – *bestehen* – Kollokator *ihr* (Autofokus)

[26] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/misc/tutorial.html übernommen [abgerufen am 5.1.2017].

KA - Lemmatisierung

Für diese Suchanfrage **(***&bestehen*) wurde zum Schluss noch die letzte KA durchgeführt, diesmal mit der Verwendung von Lemmatisierung. Die Funktionswörter wurden nicht ignoriert und der Parameter Autofokus wurde nicht ausgewählt (Abb. 17).

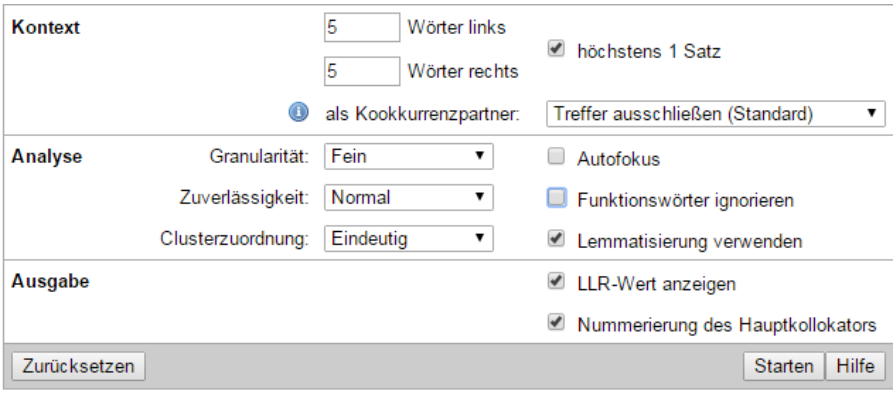


Abb. 17 – Eistellung der KA – mit *Lemmatisierung*

Falls man die Lemmatisierung ankreuzt, sollten die Kontextwörter durch ihr Lemma[27] ersetzt werden. Wichtig ist, dass diese Option automatisch die Berechnung und die Anzeige der syntagmatischen Muster ausschaltet, da eine Berechnung der syntagmatischen Muster über lemmatisierte Wortformen zur Zeit nicht definiert ist.[28] Die KA soll zeigen, dass die Wortformen in der Liste nicht einzeln ausgewertet werden, sondern alle Wortformen, die auf dasselbe Lemma zurückgeführt werden können, werden zusammen, als eine Einheit, ausgewertet (vgl. PERKUHN, 2012, 118).

Die Analyse konnte aber diese Aussage nicht überprüfen, weil sie auch nach mehreren Versuchen fehlgeschlagen ist. Auch wenn die Einstellungen geändert wurden, und die Funktionswörter aus der Analyse ausgeschlossen waren, ist die Analyse leider nicht gelungen und ist immer nach einigen Minuten im Punkt „testing statistics“ stehen geblieben (Abb. 18).

[27] Die Grundform der Kollokatoren (Infinitiv, Nominativ Singular usw.).
[28] Die Informationen wurden aus http://www.ids-mannheim.de/cosmas2/win-app/hilfe/menue/optionen/KOOK_SONSTIGE.html übernommen [abgerufen am 5.1.2017].

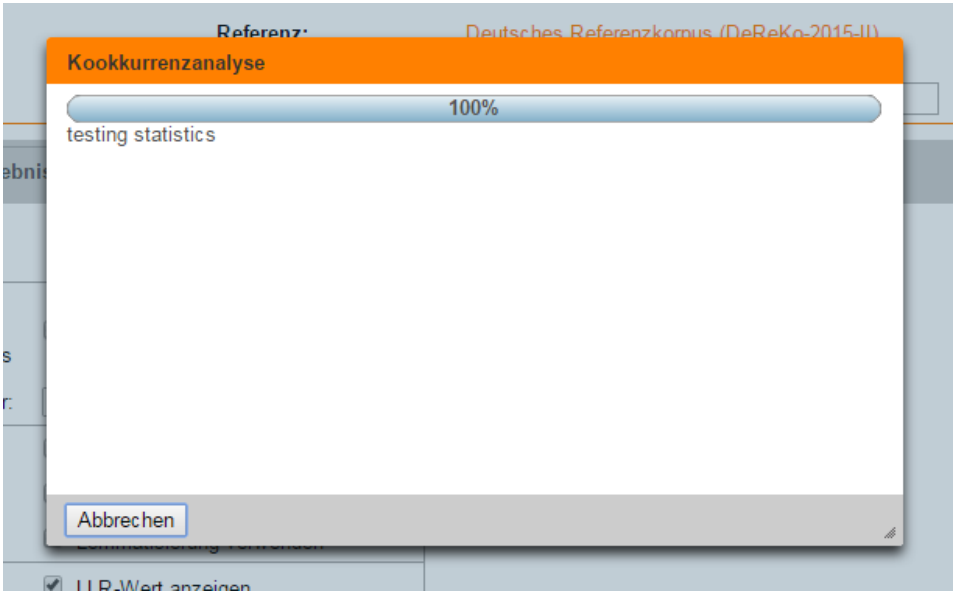


Abb. 18 – KA – Fehler

KA - Identifikation von festen Wortverbindungen

Für die nächste KA wurde das Adjektiv *schwarz* ausgewählt. Dank dieser Analyse sollte überprüft werden, inwieweit (oder ob überhaupt) die Analyse zur Identifikation von Kollokationen und anderen festen Wendungen (z.B. Phraseologismen) geeignet sein kann. Im Rahmen dieser Analyse wurden die Funktionswörter ignoriert und die Parameter Autofokus und Lemmatisierung wurden nicht verwendet.

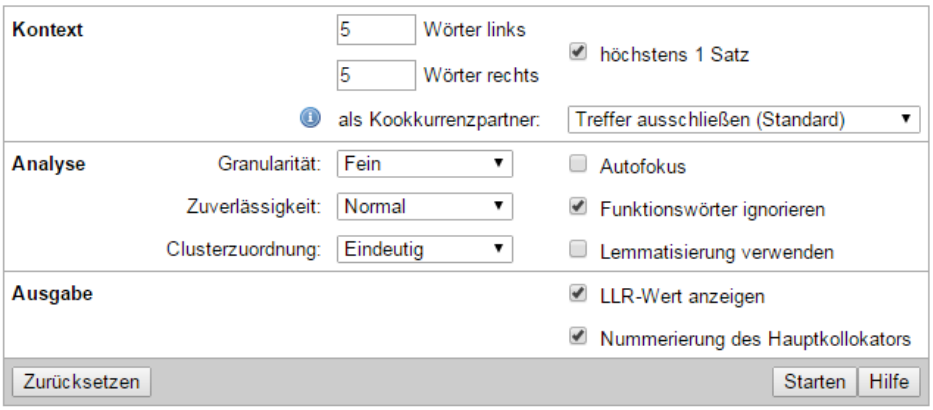


Abb. 19 – Einstellung der KA – das Adjektiv *schwarz*

Nach der Durchführung dieser Analyse ist ganz klar, dass die Analyse sehr nützlich für die Bestimmung der festen Wendungen sein kann. Gleich auf den ersten drei Plätzen in der KA lassen sich die Phraseologismen finden (Abb. 20, 21) – auf dem ersten Platz steht *Zahlen* (mit LLR 146207), auf dem zweiten Platz steht *Schafe* (mit LLR 122904) und auf dem dritten Platz steht *Meer* (mit LLR 85900). Auf den anderen Plätzen befinden sich die Phraseologismen nicht mehr (Platz 4 – 78866 *Weiß*, Platz 5 – LLR 65967 *Dr*, Platz 6 – LLR 54389 *weißen*).

⊞	1	146207	3	3	**Zahlen** schreiben schreibt	66% schwarze Zahlen [...] schreiben [...] schreibt
⊞			4	1	Zahlen schreiben operativ Verlustjahren erstmals	100% Verlustjahren erstmals ... operativ schwarze Zahlen schreiben
⊞			5	1	Zahlen schreiben operativ Verlustjahren	100% Verlustjahren ... operativ schwarze Zahlen schreiben

Abb. 20 –Ergebnis der KA – *schwarz*– Kollokator *Zahlen*

⊞	2	122904	31529	5	**Schafe** Branche überall	60% Branche wie überall [so genannte] schwarze Schafe
⊞			31542	13	Schafe Branche tummeln	38% der Branche tummeln sich auch schwarze Schafe
⊞			32303	761	Schafe Branche	35% schwarze Schafe in der Branche
⊞			32304	1	Schafe Herde Reihen	100% Herde schwarzer Schafe ... Reihen
⊞			32306	2	Schafe Herde tummeln	50% schwarze Schafe ... Herde tummeln
⊞			32397	91	Schafe Herde	28% eine Herde [...] schwarzer Schafe
⊞			32647	250	Schafe Reihen	53% gegen schwarze Schafe in den|ihren eigenen Reihen
⊞			33190	543	Schafe überall	41% Schwarze Schafe [gibt es] überall
⊞			33283	93	Schafe tummeln	61% tummeln sich viele|auch schwarze Schafe
⊞			46494	13211	Schafe	52% schwarze [...] Schafe
⊞	3	85900	46496	2	**Meer** Sotschi Krim	50% Sotschi ... Schwarzen Meer ... Krim
⊞			46527	31	Meer Sotschi Badeort	74% im russischen Badeort Sotschi am Schwarzen Meer herrschten
⊞			47028	501	Meer Sotschi	87% in Sotschi am Schwarzen Meer
⊞			47031	3	Meer Kaspischem Mittelmeer	33% Mittelmeer Schwarzem Meer ... Kaspischem Meer
⊞			47034	3	Meer Kaspischem Ostsee	66% Ostsee Schwarzem und Kaspischem Meer
⊞			47035	1	Meer Kaspischem Krim	100% Krim ... Schwarzem Meer ... Kaspischem Meer
⊞			47172	137	Meer Kaspischem	72% zwischen Schwarzem und Kaspischem Meer

Abb. 21 – Ergebnis der KA – *schwarz* – Kollokatoren *Schafe* und *Meer*

KA – Granularität

Im Rahmen dieser Analyse wurde auch der Parameter Granularität überprüft und an den Beispielen wurde gezeigt, wie sich die Analysen voneinander unterscheiden, wenn man entweder *sehr grobe* oder *feine* Granularität auswählt. Bei der feinen Granularität sind etwas unsicherere Aussagen über eventuell komplexere Kombinationen gewünscht. Die feine Granularität zielt vor allem auf die Wortverbindungen, was man auch auf dem folgenden Bildschirmfoto sehen kann (Abb.

22). Diese feine Einstellung kann zum Beispiel für die nähere Untersuchung der gefundenen Phraseologismen geeignet sein.

1	146207	3	3	**Zahlen** schreiben schreibt	66% schwarze Zahlen [...] schreiben [...] schreibt
		4	1	Zahlen schreiben operativ Verlustjahren erstmals	100% Verlustjahren erstmals ... operativ schwarze Zahlen schreiben
		5	1	Zahlen schreiben operativ Verlustjahren	100% Verlustjahren ... operativ schwarze Zahlen schreiben
		8	3	Zahlen schreiben operativ Geschäftsjahr	100% im ... Geschäftsjahr [...] operativ [...] schwarze Zahlen zu schreiben
		9	1	Zahlen schreiben operativ erstmals Quartal	100% Quartal ... erstmals operativ schwarze Zahlen ... schreiben
		22	13	Zahlen schreiben operativ erstmals	84% zweiten Halbjahr erstmals [...] operativ schwarze Zahlen schreiben
		23	1	Zahlen schreiben operativ Gesamtjahr Quartal	100% Quartal ... Gesamtjahr ... operativ schwarze Zahlen schreiben
		26	3	Zahlen schreiben operativ Gesamtjahr	100% im Gesamtjahr [...] operativ schwarze Zahlen schreiben
		30	4	Zahlen schreiben operativ Quartal	100% vierten Quartal 2003 operativ [wieder] schwarze Zahlen schreiben sagte Opel-Vorstand
		122	92	Zahlen schreiben operativ	91% operativ [wieder] schwarze Zahlen [zu] schreiben
		123	1	Zahlen schreiben schrieben	100% schrieben ... schreiben schwarze Zahlen
		129	6	Zahlen schreiben Verlustjahren erstmals	100% mehreren Verlustjahren [noch 1998] erstmals [...] schwarze Zahlen schreiben
		145	16	Zahlen schreiben Verlustjahren	100% nach zwei Verlustjahren ... wieder schwarze Zahlen schreiben
		146	1	Zahlen schreiben Mio erstmals	100% Mio ... erstmals ... schwarze Zahlen schreiben
		158	12	Zahlen schreiben Mio	83% von ... Mio ... wieder schwarze Zahlen [...] schreiben
		159	1	Zahlen schreiben schrieb	100% schwarze Zahlen schreiben schrieb
		164	5	Zahlen schreiben Geschäftsjahr erstmals	100% Geschäftsjahr [...] erstmals [...] schwarze Zahlen schreiben
		165	1	Zahlen schreiben Geschäftsjahr Quartal	100% Geschäftsjahr ... Quartal schwarze Zahlen schreiben

Abb. 22 – Ergebnis der KA (Ein Teil des Clusters) – Kollokator *Zahlen* – Granularität: fein

Im Unterschied dazu kann sehr grobe Granularität zur schnellen Übersicht über die Kollokatoren dienen. Vor allem wenn der Nutzer keine detaillierten Informationen zu den Kollokatoren braucht und nur die Liste von einzelnen Kollokatoren für ihn wichtig ist, lohnt es sich, die sehr grobe Granularität einzustellen. Auf dem folgenden Bildschirmfoto ist zu sehen, wie übersichtlich und bündig diese KA ist (Abb. 23).

⊞	1	146207	31524	31524	**Zahlen**	65% wieder schwarze [...] Zahlen schreiben
⊞	2	122904	46494	14970	**Schafe**	53% schwarze [...] Schafe
⊞	3	85900	62487	15993	**Meer**	63% am Schwarzen [...] Meer
⊞	4	78866	75692	13205	**Weiß**	76% Schwarz [auf|und] Weiß
⊞	5	65967	81188	5496	**Dr**	80% Frau Dr [...] Schwarz SPD
⊞	6	54389	91615	10427	**weißen**	31% schwarzen [und ...] weißen

Abb. 23 – Ergebnis der KA – Primärkollokatoren – Granularität: sehr grob

3.3 Beschreibung und anschauliche Durchführung der Analyse in der CCDB[29]

Wie schon erwähnt wurde, verfügt die CCDB über eine Sammlung von Ergebnissen der Kookkurrenzanalysen, die miteinander verglichen wurden. Im Jahre 2001 wurde auf der Basis von DeReKo eine Datenbank angelegt, in der die Ergebnisse von Analysen, ihre Kookkurrenzprofile zu mehr als 220.000 Grundformen mit über 770.000 verschiedenen Kookkurrenzpartnern abgelegt sind. Es werden bis zu 100.000 Verwendungen pro Wort und Analyse gespeichert.[30] Daraus ergibt sich, dass die Durchführung der Analyse für den Nutzer nicht so zeitaufwändig ist, denn die Ergebnisse der Analyse sind schon in der CCDB gespeichert und der Nutzer ruft sie nur ab.

Die Datenbank ist statisch, daraus folgt, dass die Daten nicht mehr erweitert werden, wie es im Falle der kontinuierlich erweiterten KA ist. Da die Parameter für die Durchführung der Analyse schon vordefiniert sind, bietet die CCDB dem Nutzer nicht an, sie zu ändern.

In den folgenden Absätzen wird das Suchverfahren in der CCDB anschaulich am Beispiel der Abbildungen beschrieben. Der Verlauf der Analyse ist folgend: Auf der Webseite http://corpora.ids-mannheim.de/ccdb/ gibt man das gesuchte Wort in das entsprechende Feld ein und klickt auf „anzeigen“. Schon hier beobachtet man die Unterschiede zu der KA. In der KA kann man sich entscheiden, ob man das Wort in der lemmatisierten Form eingibt oder ob man eine konkrete Wortform analysieren möchte. In der CCDB muss das Wort in der Grundform eingegeben werden und man muss dabei auf die Diakritik achten. Der nächste Unterschied ist auch das, dass an dieser Stelle in der CCDB nur Kookkurrenzprofile von einzelnen Wörtern angezeigt werden. Im Unterschied zu der KA zeigt die CCDB keine Kookkurrenzprofile von Mehrwortkombinationen.

[29] Die Informationen in diesem Kapitel wurden aus http://linguistik.zih.tu-dresden.de/lehre/blogs/blog/2015/02/04/kookkurrenzanalyse-ein-blog-ueber-die-definition-und-funktion-einer-korpuslinguistischen-methode/ übernommen [abgerufen am 8.3..2017].
[30] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 2.3..2017].

In den nächsten Absätzen wird die Analyse in der CCDB am Beispiel der Abbildungen gezeigt und beschrieben. Im Folgenden wurde das Analysewort *Katze* ausgewählt:

Abb. 24 – CCDB Startseite

Es öffnet sich ein neues Fenster – ein Kookkurrenzprofil des Wortes *Katze*:

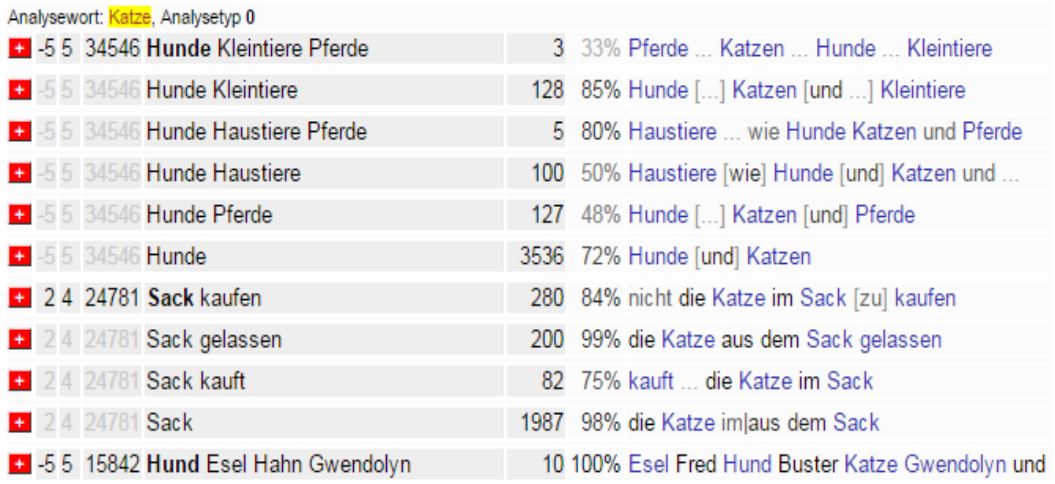


Abb. 25 – CCDB Kookkurrenzprofil

Das Kookkurrenzprofil lässt sich wie folgend entschlüsseln: Jede Konkordanz[31] ist in dieser Tabelle mit einer Kookkurrenz repräsentiert. Diese Kookkurrenzen sind durch Cluster[32] mit interner hierarchischer Struktur geordnet. Wenn man auf das „+“ am linken Rand klickt, öffnet sich die Liste der Konkordanzen für diese Kookkurrenz. Man sieht also den Kontext, in dem die Wörter in den Texten vorkommen. Die Zahlen am linken Rand - z.B. die -5 und 5 ganz oben vor dem Wort *Hunde* bedeuten, dass dieses Wort typischerweise 5 Stellen vor dem Wort *Katze* und bis zu 5 Stellen danach vorkommt[33]. Es lässt sich auch ablesen, dass das Wort *Hunde* dadurch, dass es an oberster Stelle des Kookkurrenzprofils steht, der primäre Kookkurrenzpartner von *Katze* ist. Die Zahl 34546 links neben *Hunde* zeigt den LLR-Wert (Log-Likelihood-Ratio) der Kookkurrenz von *Katze* mit dem Kookkurrenzpartner *Hunde* an. Die Zahlen in der Mitte, z.B. die 1987 rechts von *Sack,* stellen die Anzahl der Konkordanzen in diesem Cluster dar. In diesem Fall kommen *Sack* und *Katze* 1987mal zusammen im DeReKo vor. Die Prozentwerte weisen auf das dominante syntagmatische Muster[34] dieser Kookkurrenz hin. 98% der Vorkommen folgen dem Muster *die Katze im/aus dem Sack*.

Die Dechiffrierung eines Kookkurrenzprofils ist für die Zwecke dieser Diplomarbeit besonders wichtig, weil gerade die Kookkurrenzprofile der ausgewählten Wörter im praktischen Teil der Arbeit mit den Kookkurrenzprofilen der KA des IDS untereinander verglichen und bewertet werden.

Die Kookkurrenzdatenbank stellt noch viele andere Methoden der Analyse zur Verfügung (z.B. Strukturierung der Ähnlichkeitsrelationen - Selbstorganisierende semantische Karten bei der Homonymie/Polysemie Bestimmung, Vergleich von Quasi-Synonymen usw.). Die Forschung dieser Methoden betrifft aber nicht das Ziel dieser Diplomarbeit und aus dem Grund werden in der vorliegenden Arbeit diese Applikationen nicht näher angegangen.

---

[31]Die **Konkordanz**, oder auch KWIC-Darstellung, ist eine zeilenweise Auflistung von Treffern in ihrem Kontext. (Die Definition wurde aus http://www.ids-mannheim.de/cosmas2/web-app/hilfe/seite/kwic/ übernommen [abgerufen am 5.1.2017].)

[32]Die KA fasst Gruppen von Texten zusammen, die eine bestimmte Menge von Wörtern innerhalb eines Textfensters gemeinsam haben, diese Gruppen werden als (Text-)**Cluster** genannt (vgl. PERKUHN, 2012, 119).

[33] Die typische Stellung der Kollokatoren im Kontext lässt sich auch von der KA ablesen, wenn man die KA mit dem Parameter Autofokus durchführt (vgl. 3.2).

[34]Das syntagmatische Muster erfasst typische Realisierungen (lexikalische Umgebung) sowie syntaktische Struktur (vgl. PERKUHN, 2012, 122).

4 Vergleich der Ergebnisse der Kookkurrenzanalyse und der Kookkurrenzdatenbank (CCDB)

Der letze Teil der Diplomarbeit widmet sich der praktischen und anschaulichen Durchführung der beiden Analysen am Beispiel ausgewählter Wörter. Dieses Kapitel setzt sich zum Ziel, die KA und die CCDB zu vergleichen und zu überprüfen, wie (und ob überhaupt) die kontinuierliche Erweiterung eines Korpus für die Analyse wichtig ist und inwieweit die ständige Erweiterung des Korpus die Ergebnisse beeinflusst und ändert.

Die Verfasserin dieser Abreit ist der Meinung, dass die Textmenge, die in einem Korpus enthalten ist, auch einen Einfluss auf die Ergebnisse der Kookkurrenzanalysen hat. Aus diesem Grund setzt die Verfasserin voraus, dass die Ergebnisse der CCDB und der KA Differenzen aufweisen können. Im Rahmen dieses Kapitels soll die Hypothese der Verfasserin bestätigt oder widerlegt werden.

Im Rahmen dieses Kapitels wurden zehn zufällig ausgewählte Wörter je eine Wortart sowohl in der CCDB als auch in der KA analysiert. Für jedes analysierte Wort wurden die Konkordanzen erstellt, Wortliste mit Frequenzangaben erzeugt und Kookkurrenzen berechnet. Um auf die Gemeinsamkeiten und Differenzen zwischen der CCDB und der KA aufmerksam zu machen, wurden diese Angaben in jedem Unterkapitel untereinander verglichen.

Die CCDB ist statisch, daraus folgt, dass die Daten nicht mehr erweitert werden, wie es im Falle der kontinuierlich erweiterten KA ist. Da die Parameter für die Durchführung der Analyse in der CCDB schon vordefiniert sind, bietet die Datenbank dem Nutzer keine Möglichkeit, sie zu ändern. Das einzige, was der Nutzer beeinflussen kann, ist die Wahl *mit* oder *ohne Synsemantika*[35]. Wenn die Synsemantika in die CCDB-Analyse im Rahmen dieser Diplomarbeit einbezogen wurden, wurde auch die KA parallel dazu mit Synsemantika durchgeführt. Im Unterschied zu der CCDB lassen sich auch andere Parameter in der KA beliebig ändern (vgl. 3.1), was auch die Ergebnisse der Analyse beeinflussen kann. Im Rahmen der einzelnen Analysen wurde

[35] Das **Synsemantikon** (auch **Funktionswort**) ist ein inhaltsarmes Wort, das seine eigentliche Bedeutung erst durch den umgebenden Text erhält. (Die Definition wurde aus http://www.duden.de/suchen/dudenonline/synsemantikon übernommen [abgerufen am 20.6..2017].)

die Grundeinstellung der KA in begründeten Fällen geändert (z.B. *Granularität: sehr grob* hat schnellere und übersichtlichere Orientierung unter den Primärkollokatoren ermöglicht).

Für die Zwecke der Recherche wurden sowohl flektierbare als auch nicht flektierbare Wortarten ausgewählt. Es handelt sich um folgende Wörter:

- **flektierbare** Wortarten
  - **deklinierbare** Wortarten
  - das Substantiv *Gefahr*
  - das Adjektiv *schwarz*
  - der Artikel *der*
  - das Pronomen *man*
  - das Numerale *zweierlei*

  - **konjugierbare** Wortart
  - das Verb *bestehen*

- **nicht flektierbare** Wortarten
  - das Adverb *damals*
  - die Präposition *angesichts*
  - die Konjunktion *wenngleich*
  - die Interjektion *ach*

4.1 Analyse eines Substantivs

Für die erste Analyse wurde das Substantiv ***Gefahr*** ausgewählt. Es soll gezeigt werden, ob die beiden Analysen (die KA und die Analyse in der CCDB) gleiche Hinweise für gemeinsames Auftreten von Wörtern (Kookkurrenzen) und ein gleiches Maß für deren Affinität (Kohäsion) liefern.

Zunächst wurde die KA des IDS durchgeführt. Nach der Auswahl der Korpora (W-öffentlich – alle öffentlichen Korpora des Archivs W) gibt man eine Suchanfrage ins System ein. In diesem Fall wurde eine simple Abfrage nach dem Wort *Gefahr,* in seiner lemmatisierten Form *&Gefahr,* formuliert. Nach der Bewilligung aller Wortformen (die kleinen und großen Buchstaben usw.) werden die Ergebnisse geladen (703 409 Treffer), die Ergebnisliste wird angezeigt und im WWW-Client wird der Reiter KA gewählt.

In dieser KA sind die vordefinierten Einstellungen *Zuverlässigkeit: normal* und *Clusterzuordnung: eindeutig* ohne Änderung geblieben, nur das Maß der Granularität wurde auf *sehr grob* geändert. Im Unterschied zur *feinen Granularität* kann *sehr grobe Granularität* zur schnellen Übersicht über die Kollokatoren dienen, was für diese Recherche geeignet ist. Diese Analyse ist sehr übersichtlich und bündig, die sekundären (evtl. tertiären) Kollokatoren werden hier nicht gezeigt. In diesem Fall wird *sehr grobe Granularität* als eine ideale Form für den Vergleich der Primärkollokatoren mit der Analyse der CCDB dienen.

Im Bereich *Kontext definieren* wurden auch keine Änderungen vorgenommen und es bleiben hier 5 Wörter rechts und 5 Wörter links bewahrt. Funktionswörter wurden in dieser KA ignoriert und die Parameter *Autofokus* und *Lemmatisierung verwenden* wurden nicht ausgewählt.

Als Ergebnis dieser KA bekommt man eine Liste von Wörtern aus der Umgebung des Bezugswortes (*Gefahr*), mit ihrer Häufigkeit und ihrer statistischen Bewertung. Mit dem Symbol „#“ werden die primären, in fetter Schrift wiedergegebenen, Partnerwörter nummeriert. Für den Vergleich mit der Analyse in CCDB wurden die ersten 15 Primärkollokatoren ausgewählt.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	213082	39272	39272	**besteht**	69% Es besteht [die] Gefahr dass ...
⊞	2	108599	53338	14066	**bestehe**	78% Es bestehe [die] Gefahr dass ...
⊞	3	78053	62754	9416	**birgt**	45% birgt [... die] Gefahr dass ...
⊞	4	60225	68651	5897	**Verzug**	99% Gefahr im|in Verzug
⊞	5	48100	74408	5757	**gebannt**	94% die Gefahr [... nicht] gebannt
⊞	6	47171	84515	10107	**droht**	52% Gefahr [...] droht
⊞	7	46412	96837	12322	**läuft**	79% läuft [...] Gefahr
⊞	8	40127	106905	10068	**laufen**	55% laufen [...] Gefahr
⊞	9	33105	110753	3848	**lauern**	51% Gefahren [die ...] lauern
⊞	10	33033	117302	6549	**bestand**	56% Es bestand [die|keine] Gefahr dass die
⊞	11	26155	129979	12677	**sieht**	66% sieht [... die in] Gefahr
⊞	12	25609	138538	8559	**groß**	68% ist die|Die Gefahr [ist] groß dass
⊞	13	23805	140558	2020	**akute**	94% eine|keine akute [...] Gefahr für ...
⊞	14	22039	143935	3377	**Leib**	78% Gefahr für Leib und Leben
⊞	15	20277	146240	2305	**Besteht**	100% Besteht [nicht die] Gefahr dass ...

Abb. 26 – Ergebnis der KA – das Substantiv *Gefahr*

Wie es auf der Abbildung 26 zu sehen ist, sind 4 von 15 Primärkollokatoren die verschiedenen Formen des Verbs *bestehen* (*besteht*, *bestehe*, *bestand* und *Besteht*). Die anderen Primärkollokatoren sind sowohl die Verben als auch Substantive und Adjektive.

Wenn man die gleiche Analyse in der CCDB durchführt, gibt das System dem Nutzer fast keine Möglichkeiten, die Eistellungen zu ändern. Das einzige, was der Nutzer beeinflussen kann, ist die Wahl *mit* oder *ohne Synsemantika*. In diesem Fall wurde die Analyse *ohne Synsemantika* durchgeführt. Man muss auch bei der Formulierung der Suchanfrage auf die Großschreibung achten. Als Ergebnis dieser Analyse bekommt man eine detaillierte Liste, die dem Nutzer sowohl die Primär- als auch Sekundärkollokatoren zur Verfügung stellt.

Analysewort: Gefahr, Analysetyp 0

-4	1	28602	**besteht** Sonst	66	100%	Sonst besteht [die] Gefahr dass|daß ...
-4	1	28602	besteht sonst	72	50%	sonst besteht [die] Gefahr daß|dass ...
-4	1	28602	besteht Zudem	37	97%	Zudem besteht die Gefahr daß|dass ...
-4	1	28602	besteht	5152	69%	Es besteht [die] Gefahr dass|daß ...
-5	5	20310	**bestehe** Andernfalls	33	100%	Andernfalls bestehe [die] Gefahr daß die
-5	5	20310	bestehe Ansonsten	29	100%	Ansonsten [...] bestehe [die] Gefahr dass|daß die
-5	5	20310	bestehe Zudem	40	100%	Zudem bestehe die Gefahr dass|daß ...
-5	5	20310	bestehe	2703	78%	Es bestehe [die] Gefahr dass|daß die
2	2	11642	**Verzug** ist sei	3	66%	sei ... wenn Gefahr im Verzug ist
2	2	11642	Verzug ist	258	46%	ist [...] Gefahr im|in Verzug
2	2	11642	Verzug sei	78	50%	Gefahr in|im Verzug [...] sei
2	2	11642	Verzug sah	12	91%	sah [...] Gefahr im|in Verzug und
2	2	11642	Verzug	874	99%	Gefahr im|in Verzug
-5	3	9558	**birgt** auch gewisse	3	66%	birgt ... auch [...] gewisse Gefahren
-5	3	9558	birgt auch enorme	1	100%	birgt enorme ... auch Gefahren
-5	3	9558	birgt auch	241	49%	birgt [aber] auch [...] Gefahren
-5	3	9558	birgt gewisse	11	45%	birgt [... aber|auch] gewisse Gefahren
-5	3	9558	birgt enorme	7	71%	birgt enorme [...] Gefahren
-5	3	9558	birgt	1223	48%	birgt [... die] Gefahr

Abb. 27 – Ergebnis der Analyse in der CCDB – das Substantiv *Gefahr*

Die ersten 15 Primärkollokatoren in der CCDB sind (absteigend geordnet): *besteht, bestehe, Verzug, birgt, läuft, gebannt, größte, droht, große, laufen, akute, sieht, Leib, bestand, groß*. Wie es zu sehen ist, sind die ersten zwei Primärkollokatoren mit den Primärkollokatoren in der KA identisch – *besteht* und *bestehe*. Wenn man diese zwei Kollokatoren mit der KA vergleicht, weisen sie einen ganz anderen LLR auf (*besteht* – KA LLR 213082, CCDB LLR 28602; bestehe – KA LLR 108599, CCDB LLR 20310). Der LLR-Wert ist eine statistisch berechnete Größe der Affinität des Kookkurrenzpartners zum Bezugswort und die Zahl hängt auch mit der Korpusgröße zusammen.

Die anderen Primärkollokatoren (*Verzug, birgt, läuft, gebannt, droht, laufen, akute, sieht, Leib, bestand, lauern, groß*) sind auch identisch, nur die Reihenfolge und die Wortformen (*groß – größte* usw.) sind mit der KA nicht gleich.

4.2 Analyse eines Adjektivs

Für die nächste Analyse und für den nächsten Vergleich wurde das Adjektiv ***schwarz*** ausgewählt. Die detaillierte KA dieses Adjektivs, mit der Untersuchung der Einstellungsmöglichkeiten, wurde schon im dritten Kapitel durchgeführt (vgl. 3.2). Dank dieser Analyse wurde überprüft, dass die KA zur Identifikation von Kollokationen und anderen festen Wendungen (z.B. Phraseologismen) geeignet ist. Im Rahmen dieser Analyse soll überprüft werden, ob die KA und CCDB die gleichen Ergebnisse und gleiche phraseologische Kollokatoren und Belege liefern.

Zunächst wurde die KA des IDS durchgeführt. Im Rahmen dieser Analyse wurden die Funktionswörter ignoriert und die Parameter *Autofokus* und *Lemmatisierung* wurden nicht verwendet. Um die Ergebnisliste detaillierter sehen zu können (auch mit den Sekundär- bzw. Tertiärkollokatoren), wurde die *feine Granularität* bewahrt. Es wurde eine simple Abfrage *&schwarz* formuliert und alle Wortformen wurden erlaubt. Das System bietet 1 136 833 Ergebnisse, die Ergebnisliste wird angezeigt und im WWW-Client wird der Reiter KA gewählt.

Bei solcher Einstellung bietet die KA über mehr als 100 Kookkurrenzen zum Primärkollokator *Zahlen*.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	146207	3	3	**Zahlen** schreiben schreibt	66% schwarze Zahlen [...] schreiben [...] schreibt
⊞			4	1	Zahlen schreiben operativ Verlustjahren erstmals	100% Verlustjahren erstmals ... operativ schwarze Zahlen schreiben
⊞			5	1	Zahlen schreiben operativ Verlustjahren	100% Verlustjahren ... operativ schwarze Zahlen schreiben
⊞			8	3	Zahlen schreiben operativ Geschäftsjahr	100% im ... Geschäftsjahr [...] operativ [...] schwarze Zahlen zu schreiben
⊞			9	1	Zahlen schreiben operativ erstmals Quartal	100% Quartal ... erstmals operativ schwarze Zahlen ... schreiben
⊞			22	13	Zahlen schreiben operativ erstmals	84% zweiten Halbjahr erstmals [...] operativ schwarze Zahlen schreiben
⊞			23	1	Zahlen schreiben operativ Gesamtjahr Quartal	100% Quartal ... Gesamtjahr ... operativ schwarze Zahlen schreiben
⊞			26	3	Zahlen schreiben operativ Gesamtjahr	100% im Gesamtjahr [...] operativ schwarze Zahlen schreiben
⊞			30	4	Zahlen schreiben operativ Quartal	100% vierten Quartal 2003 operativ [wieder] schwarze Zahlen schreiben sagte Opel-Vorstand
⊞			122	92	Zahlen schreiben operativ	91% operativ [wieder] schwarze Zahlen [zu] schreiben
⊞			123	1	Zahlen schreiben schrieben	100% schrieben ... schreiben schwarze Zahlen
⊞			129	6	Zahlen schreiben Verlustjahren erstmals	100% mehreren Verlustjahren [noch 1998] erstmals [...] schwarze Zahlen schreiben
⊞			145	16	Zahlen schreiben Verlustjahren	100% nach zwei Verlustjahren ... wieder schwarze Zahlen schreiben
⊞			146	1	Zahlen schreiben Mio erstmals	100% Mio ... erstmals ... schwarze Zahlen schreiben
⊞			158	12	Zahlen schreiben Mio	83% von ... Mio ... wieder schwarze Zahlen [...] schreiben

Abb. 28 – Ergebnis der KA (Ein Teil des Clusters) – Kollokator *Zahlen* – Granularität: fein

Schon auf den ersten Blick ist es zu sehen, dass die CCDB eine visuell ganz unterschiedliche Ergebnistabelle liefert.

1	1	55563	**Zahlen** schreiben schreibt	1	100%	schwarze Zahlen ... schreiben schreibt
1	1	55563	Zahlen schreiben	1334	87%	wieder schwarze [...] Zahlen [...] schreiben
1	1	55563	Zahlen schreibt	624	65%	schreibt [... wieder] schwarze Zahlen
1	1	55563	Zahlen schreibe	134	57%	schreibe [...] schwarze [...] Zahlen
1	1	55563	Zahlen	5230	72%	wieder schwarze [...] Zahlen schreiben
1	1	25412	**Schafe**	1939	63%	schwarze [...] Schafe
2	2	11702	**weiß** gelb grau	2	50%	gelb ... schwarz ... weiß grau
2	2	11702	weiß gelb Farben	3	33%	Farben ... weiß ... schwarz ... gelb
2	2	11702	weiß gelb	38	26%	rot weiß [...] schwarz [und] gelb
2	2	11702	weiß grau Farben	2	100%	Farben [...] schwarz weiß grau
2	2	11702	weiß grau	34	44%	schwarz [...] weiß [...] grau und
2	2	11702	weiß Farben	32	71%	in den Farben [...] schwarz [und] weiß und
2	2	11702	weiß	2537	77%	schwarz [auf] weiß
-5	5	9617	**weißen** Hosen Hemden	16	50%	in schwarzen Hosen [und] weißen Hemden
-5	5	9617	weißen Hosen	37	32%	in weißen Hemden|Trikots und schwarzen [...] Hosen
-5	5	9617	weißen Hemden	38	50%	in|mit schwarzen Hosen|Anzügen und weißen [...] Hemden und
-5	5	9617	weißen Tasten	48	81%	Auf|die schwarzen und auf weißen Tasten
-5	5	9617	weißen	1687	38%	schwarzen [und ...] weißen
1	1	9562	**Loch**	961	51%	in ein schwarzes [...] Loch
2	3	9547	**schreiben** Null erstmals	1	100%	erstmals ... schwarze Null ... schreiben

Abb. 29 – Ergebnis der CCDB – das Adjektiv *schwarz*

Die KA liefert detailliertere Aufschlüsselung der Kollokatoren und der Nutzer findet hier im Unterschied zur CCDB viel detailliertere Cluster (Abb. 28). Die CCDB liefert die Kollokatoren nur im kleineren Maß als die KA. Für detaillierte Untersuchung der Verwendung eines Phraseologismus scheint also die KA geeigneter zu sein, der Nutzer bekommt in der KA mehrere Beispiele, aus denen er die präziseren Hypothesen von der Verwendung der Phraseologismen ableiten kann.

Wie es auf der Abbildung 28 zu sehen ist, liefert die KA dank der *feinen Granularität* neben den Sekundärkollokatoren auch die Tertiärkollokatoren, die die CCDB nicht zur Verfügung stellt. Nach der Untersuchung der Sekundärkollokatoren, die bei dem Primärkollokator *Zahlen* stehen, lässt sich behaupten, dass sie in beiden Analysen identisch sind.

Falls man eine schnelle Übersicht über die häufigsten Kollokatoren zu dem Adjektiv *schwarz* gewinnen will, lohnt es sich, die Einstellung in der KA auf *grobe Granularität* zu ändern. Der Nutzer entscheidet also, wie intensiv die Mehrwortgruppen gesucht werden sollen, d.h. wie viele der nach Signifikanz sortierten

Kookkurrenzpartner als möglicher Kandidat eines Kookkurrenzpartners in Frage kommen. In der folgenden Analyse sieht es praktisch so aus, dass der Nutzer auswählt, dass lieber sichere Aussagen über kurze Kombinationen gewünscht sind. Um die Kookkurrenzen in der KA und in der CCDB vergleichen zu können, wurden wieder die ersten 15 häufigsten Kollokatoren ausgewählt.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	146207	31524	31524	**Zahlen**	65% wieder schwarze [...] Zahlen schreiben
⊞	2	122904	46494	14970	**Schafe**	53% schwarze [...] Schafe
⊞	3	85900	62487	15993	**Meer**	63% am Schwarzen [...] Meer
⊞	4	78866	75692	13205	**Weiß**	76% Schwarz [auf|und] Weiß
⊞	5	65967	81188	5496	**Dr**	80% Frau Dr [...] Schwarz SPD
⊞	6	54389	91615	10427	**weißen**	31% schwarzen [und ...] weißen
⊞	7	53212	101015	9400	**Loch**	31% ein schwarzes [...] Loch
⊞	8	49150	109820	8805	**Haare**	61% hat|kurze schwarze [...] Haare und
⊞	9	46805	117576	7756	**Null**	68% eine schwarze [...] Null
⊞	10	46041	124703	7127	**Alice**	97% Alice [...] Schwarzer
⊞	11	41753	130462	5759	**Hans-Werner**	99% Vizepräsident Hans-Werner [...] Schwarz
⊞	12	41062	148551	18089	**weiß**	68% schwarz [auf] weiß
⊞	13	40727	154539	5988	**Hose**	37% eine schwarze [...] Hose und
⊞	14	38139	161609	7070	**weiße**	27% schwarze [und] weiße
⊞	15	35163	167976	6367	**trug**	51% Er trug [... eine] schwarze ... und

Abb. 30 – Ergebnis der KA – das Adjektiv *schwarz* – Granularität: grob

Die ersten 15 Primärkollokatoren in der KA sind (absteigend geordnet): *Zahlen, Schafe, Meer, Weiß, Dr, weißen, Loch, Haare, Null, Alice, Hans-Werner, weiß, Hose, weiße, trug*. Nach der gründlicheren Untersuchung aller Kollokatoren und nach dem Anklicken der KWICs sieht man, dass es sich bei mehreren Kollokatoren um die Phraseologismen handelt (*Zahlen, Schafe, Meer, Loch, Null*).

Die ersten 15 Primärkollokatoren der CCDB Analyse sind (absteigend geordnet): *Zahlen, Schafe, weiß, weißen, Loch, schreiben, Null, Haare, Liste, Humor, Anzug, weiße, Schaf, Kassen, Hose.* Im Unterschied zur KA befinden sich in der CCDB unter den ersten 15 Kollokatoren auch die Substantive *Liste* (LLR 7009) und *Humor* (LLR 6775), also Kollokatoren, die zusammen mit dem Adjektiv *schwarz* eine phraseologische Bedeutung tragen. Diese Kollokatoren stehen in der KA erst auf dem Platz 18 (Humor – LLR 30749) und 30 (Liste – LLR 21747). Die anderen vier Primärkollokatoren mit der phraseologischen Bedeutung (*Zahlen, Schafe, Loch, Null*) lassen sich sowohl unter den ersten 15 Primärkollokatoren der KA, als auch unter den

ersten 15 Primärkollokatoren in der CCDB finden. Im Unterschied dazu lässt sich aber der Kollokator *Meer* in der CCDB nicht finden, was von dem unterschiedlichen Textmaterial beider Analysenmethoden zeugt.

4.3 Analyse eines Artikels

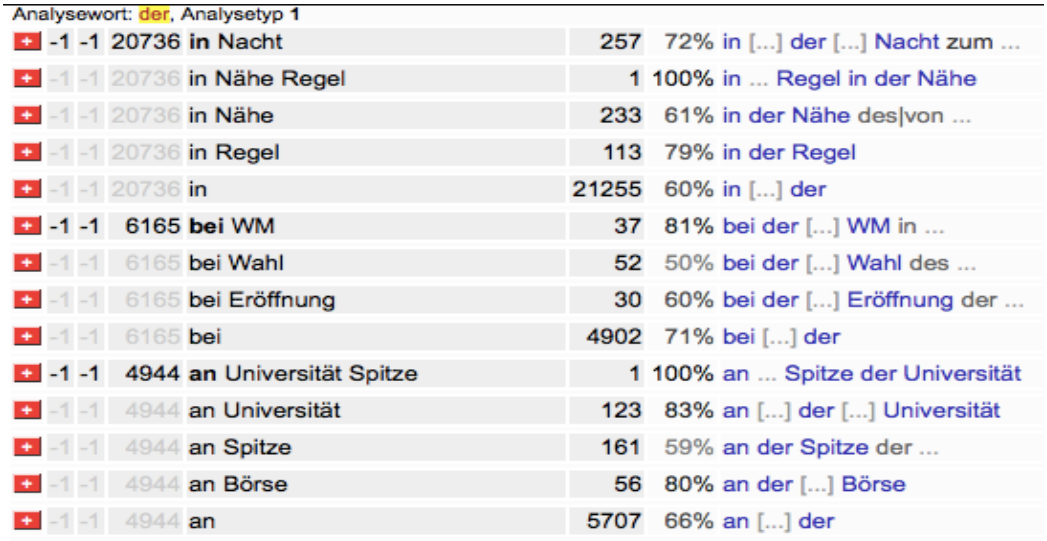
Die nächste Analyse betrifft den Artikel, also eine Wortart, die keine inhaltliche Bedeutung hat. Für die Zwecke dieser Analyse wurde der unbestimmte Artikel ***der*** ausgewählt. Da der Artikel ein grammatisches Wort ist, das regelmäßig ein Substantiv begleitet, wurden in dieser Analyse nicht die Kollokatoren untereinander verglichen, sondern es wurde hier v. a. die formale Seite des ersten Ergebnisses sowohl in der CCDB als auch in der KA analysiert und beschrieben.

In der KA wurden die Einstellungen nicht geändert (d.h. die Einstellung *feine Granularität* bleibt) und die Funktionswörter wurden in diesem Fall von der Analyse nicht ausgeschlossen. Das Korpus liefert für den Artikel *der* eine riesige Menge von Treffern – genau 231 097 189. Für die KA wählt das System automatisch nur 10 000 000 Treffer aus und es handelt sich dabei um eine Zufallsauswahl des Systems. Da es sich um eine so umfassende KA handelt, kann sie auch mehr als zwei Stunden nehmen.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	76732	2043148	2043148	in	60% in [...] der
⊞	2	61041	2043164	16	**bzw** Zufluss	100% langer linker|rechter bzw ... Zufluss der ... im
⊞			2050897	7733	bzw	49% der [...] bzw
⊞	3	43256	2050902	5	**Dr** Prof Abteilung Chefarzt	80% Prof Dr ... Chefarzt der Inneren Abteilung
⊞			2050906	4	Dr Prof Abteilung	50% Prof Dr ... der Abteilung
⊞			2050915	9	Dr Prof Abg	66% der [...] Abg Prof Dr ... Porsch
⊞			2050916	1	Dr Prof Klinik Chefarzt Neurologischen	100% Chefarzt der Neurologischen Klinik Prof Dr
⊞			2050924	8	Dr Prof Klinik Chefarzt	100% Prof Dr ... Chefarzt der [Medizinischen] Klinik für ...
⊞			2050926	2	Dr Prof Klinik Neurologischen	100% Prof Armin Grau Direktor der Neurologischen Klinik [...] Dr
⊞			2050934	8	Dr Prof Klinik	75% Prof Dr ... Direktor der [...] Klinik für ...
⊞			2050935	1	Dr Prof Professor	100% Prof Dr ... Professor der
⊞			2050936	1	Dr Prof Chefarzt Chirurgie	100% Prof Dr ... Chefarzt der Chirurgie

Abb. 31 – Ergebnis der KA – der Artikel *der*

Die gleiche Analyse des Artikels *der* wurde auch in der CCDB durchgeführt, in diesem Fall wurden die Synsemantika auch ein Bestandteil der Analyse.



Analysewort: der, Analysetyp 1

-1	-1	20736	**in** Nacht	257	72%	in [...] der [...] Nacht zum ...
-1	-1	20736	in Nähe Regel	1	100%	in ... Regel in der Nähe
-1	-1	20736	in Nähe	233	61%	in der Nähe des|von ...
-1	-1	20736	in Regel	113	79%	in der Regel
-1	-1	20736	in	21255	60%	in [...] der
-1	-1	6165	**bei** WM	37	81%	bei der [...] WM in ...
-1	-1	6165	bei Wahl	52	50%	bei der [...] Wahl des ...
-1	-1	6165	bei Eröffnung	30	60%	bei der [...] Eröffnung der ...
-1	-1	6165	bei	4902	71%	bei [...] der
-1	-1	4944	**an** Universität Spitze	1	100%	an ... Spitze der Universität
-1	-1	4944	an Universität	123	83%	an [...] der [...] Universität
-1	-1	4944	an Spitze	161	59%	an der Spitze der ...
-1	-1	4944	an Börse	56	80%	an der [...] Börse
-1	-1	4944	an	5707	66%	an [...] der

Abb. 32 – Ergebnis der CCDB – der Artikel *der*

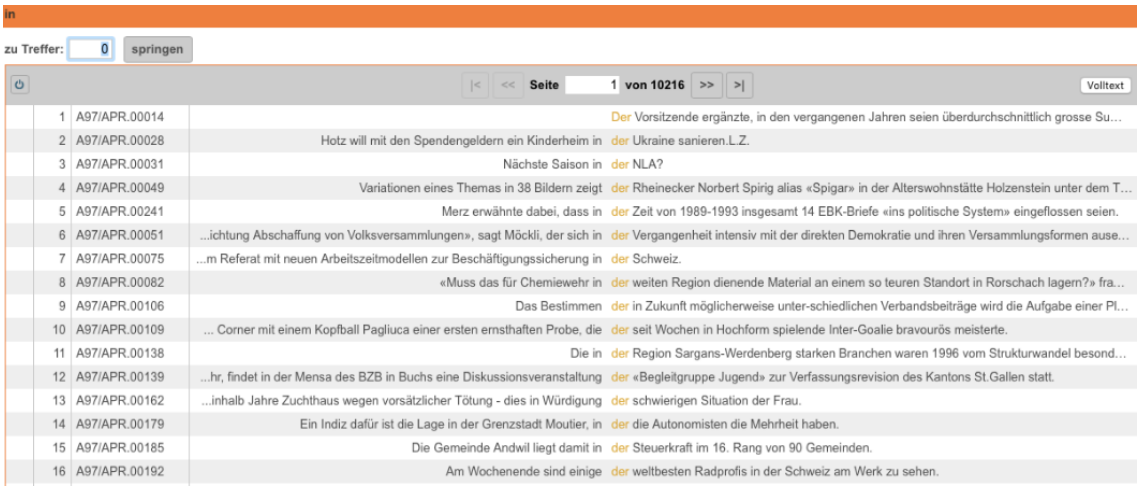
Wenn man die beiden Ergebnistabellen anschaut, ist das schon auf den ersten Blick klar, dass sich die Ergebnisse voneinander unterscheiden. Den Platz 1 - die Position des häufigsten Primärkollokators - besitzt in beiden Analysen die Präposition *in*. Als zweiter Primärkollokator erscheint in der KA die Abkürzung *bzw* und als dritter die Abkürzung *Dr*. Diese beiden Abkürzungen lassen sich in der Liste der CCDB Analyse gar nicht finden, die CCDB bewertet diese Wörter als *statistisch unspezifisch*. Diese Beispiele können belegen, wie wichtige Rolle die im Korpus beinhaltende Textauswahl spielt und wie gründlich sie die Ergebnisse ändern kann. Während sich das Korpus, aus dem die KA die Ergebnisse herleitet, ständig verbreitet und innoviert, bleibt die CCDB unveränderlich stehen. Daraus folgt, dass die CCDB auch keine Entwicklung der Sprache widerspiegeln kann. Im Falle des Artikels *der* handelt es sich um eine so frequentierte Wortart, dass am Anfang der Analyse auch keine These stehen konnte, dass die Ergebnisse gleich sein werden. Der Fakt, dass der zweit- und dritthäufigste KA-Kollokator gar nicht in der Liste der Kollokatoren der CCDB zu finden ist, zeugt davon, dass sich die beiden Textsammlungen voneinander abheben und dass sie dem Nutzer auch ganz unterschiedliche Ergebnisse liefern können.

Wenn man nur den ersten Primärkollokator *in* detaillierter untersucht, kann man auch Unterschiede beobachten (Abb. 31, 32).

Wie es auf der Abbildung 31 deutlich zu sehen ist, liefert die KA nur das primäre Partnerwort *in*. In der Nebenspalte sieht man das typische syntagmatische Muster, in dem die Wörter zusammen auftreten (hier *in […] der*). Die Prozentangabe in dieser Spalte bezieht sich auf die Häufigkeit der Reihenfolge der konkreten Wortformen, die bei der Kookkurrenzanalyse beteiligt waren (vgl. PERKUHN, 2012, 122). Allgemein lässt sich sagen, dass diese Spalte typische Konstruktionen der Kookkurrenzen zeigt. Bezogen auf die Anzahl der Textstellen wird in 60% der Fälle das Wort *in* eine Position vor *der* beobachtet. Die eckigen Klammern des Ausdrucks zwischen *in* und *der* deuten an, dass die beiden Wörter nicht zwingend unmittelbar nacheinander folgen müssen und dass es sich auch optional ein Ausdruck dazwischen schieben kann.

Im Falle der Analyse in der CCDB bekommt man nicht nur das primäre Partnerwort, sondern auch das sekundäre und tertiäre (Abb. 32) – also eine Cluster mit Kookkurrenzen. Genauso wie im Falle der KA bietet auch die CCDB die dominanten syntagmatischen Muster mit Abdeckungsrate (hier 72% *in […]der[…]Nacht zum;* 100% *in … Regel in der Nähe;* 61% *in der Nähe des/von;* 79% *in der Regel;* 60% *in […] der*). Das letztgenannte syntagmatische Muster ist mit dem syntagmatischen Muster der KA identisch und auch die Prozentangaben sind gleich.

Nach dem Anklicken des Symbols „+“ (sowohl in der KA als auch in der CCDB) liefert das System eine KWIC-Ansicht (Abb. 33, 34). Die konkreten Textbelege sind in beiden Analysen ganz unterschiedlich, was auch mit der riesigen Erweiterung der IDS-Korpora zusammenhängt. Für die KA wurden zufällig nur 10 000 000 Treffer ausgewählt, es lässt sich also erwarten, dass die Textbelege bei jeder durchgeführten KA unterschiedlich sein können.



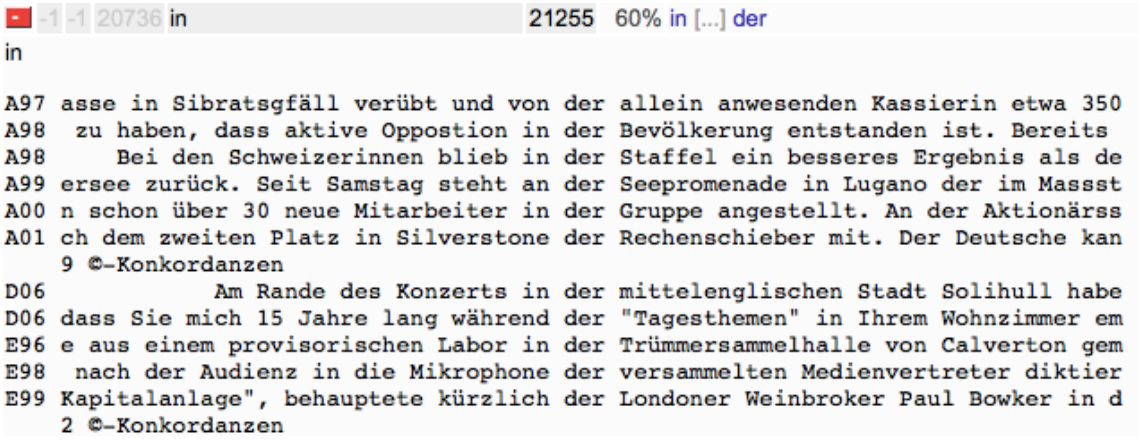
in

zu Treffer: 0 springen

|< << Seite 1 von 10216 >> >| Volltext

1	A97/APR.00014		Der Vorsitzende ergänzte, in den vergangenen Jahren seien überdurchschnittlich grosse Su…
2	A97/APR.00028	Hotz will mit den Spendengeldern ein Kinderheim in	der Ukraine sanieren.L.Z.
3	A97/APR.00031	Nächste Saison in	der NLA?
4	A97/APR.00049	Variationen eines Themas in 38 Bildern zeigt	der Rheinecker Norbert Spirig alias «Spigar» in der Alterswohnstätte Holzenstein unter dem T…
5	A97/APR.00241	Merz erwähnte dabei, dass in	der Zeit von 1989-1993 insgesamt 14 EBK-Briefe «ins politische System» eingeflossen seien.
6	A97/APR.00051	…ichtung Abschaffung von Volksversammlungen», sagt Möckli, der sich in	der Vergangenheit intensiv mit der direkten Demokratie und ihren Versammlungsformen ause…
7	A97/APR.00075	…m Referat mit neuen Arbeitszeitmodellen zur Beschäftigungssicherung in	der Schweiz.
8	A97/APR.00082	«Muss das für Chemiewehr in	der weiten Region dienende Material an einem so teuren Standort in Rorschach lagern?» fra…
9	A97/APR.00106	Das Bestimmen	der in Zukunft möglicherweise unter-schiedlichen Verbandsbeiträge wird die Aufgabe einer Pl…
10	A97/APR.00109	… Corner mit einem Kopfball Pagliuca einer ersten ernsthaften Probe, die	der seit Wochen in Hochform spielende Inter-Goalie bravourös meisterte.
11	A97/APR.00138	Die in	der Region Sargans-Werdenberg starken Branchen waren 1996 vom Strukturwandel besond…
12	A97/APR.00139	…hr, findet in der Mensa des BZB in Buchs eine Diskussionsveranstaltung	der «Begleitgruppe Jugend» zur Verfassungsrevision des Kantons St.Gallen statt.
13	A97/APR.00162	…inhalb Jahre Zuchthaus wegen vorsätzlicher Tötung - dies in Würdigung	der schwierigen Situation der Frau.
14	A97/APR.00179	Ein Indiz dafür ist die Lage in der Grenzstadt Moutier, in	der die Autonomisten die Mehrheit haben.
15	A97/APR.00185	Die Gemeinde Andwil liegt damit in	der Steuerkraft im 16. Rang von 90 Gemeinden.
16	A97/APR.00192	Am Wochenende sind einige	der weltbesten Radprofis in der Schweiz am Werk zu sehen.

Abb. 33 – KWICs der KA – der Artikel *der*



-1 -1 20736 in 21255 60% in […] der

in

```
A97 asse in Sibratsgfäll verübt und von der allein anwesenden Kassierin etwa 350
A98  zu haben, dass aktive Oppostion in der Bevölkerung entstanden ist. Bereits
A98     Bei den Schweizerinnen blieb in der Staffel ein besseres Ergebnis als de
A99 ersee zurück. Seit Samstag steht an der Seepromenade in Lugano der im Massst
A00 n schon über 30 neue Mitarbeiter in der Gruppe angestellt. An der Aktionärss
A01 ch dem zweiten Platz in Silverstone der Rechenschieber mit. Der Deutsche kan
    9 ©-Konkordanzen
D06            Am Rande des Konzerts in der mittelenglischen Stadt Solihull habe
D06 dass Sie mich 15 Jahre lang während der "Tagesthemen" in Ihrem Wohnzimmer em
E96 e aus einem provisorischen Labor in der Trümmersammelhalle von Calverton gem
E98  nach der Audienz in die Mikrophone der versammelten Medienvertreter diktier
E99 Kapitalanlage", behauptete kürzlich der Londoner Weinbroker Paul Bowker in d
    2 ©-Konkordanzen
```

Abb. 34 - KWICs der CCDB – der Artikel *der*

4.4 Analyse eines Pronomens

Die folgende Analyse betrifft das Indefinitpronomen ***man***. Diese Analyse, die wieder in der KA und in der CCDB durchgeführt wird, soll zur Feststellung von Vorkommenskorrelationen dienen, insbesondere zeigt sie, ob das Indefinitpronomen *man* mit einem anderen Wort häufiger zusammen vorkommt als es eine Zufallsverteilung von *man* und einem anderen Wort erwarten ließe.

In der KA wurden die Funktionswörter von der Analyse ausgeschlossen und die Einstellung der Grundparameter in der KA blieb ohne Änderung. Da es sich wieder um ein hochfrequentiertes Wort handelt, wählt das System für die KA 10 000 000 aus 14 403 256 Treffern zufällig aus. Diese enorme Menge von Treffern verursacht, dass der Verlauf der KA mehr als eine Stunde dauern kann.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	365406	3	3	**sieht** hinschaut Überall	66% Überall wo man hinschaut sieht man
⊞			313	310	sieht hinschaut	45% Wenn|wenn man [genau|genauer] hinschaut [...] sieht man dass
⊞			314	1	sieht förmlich lauter	100% man sieht förmlich lauter
⊞			315	1	sieht förmlich Gesichtern	100% Man sieht förmlich ... Gesichtern
⊞			606	291	sieht förmlich	51% Man [...] sieht [...] förmlich wie
⊞			607	1	sieht umschaut Gesichtern	100% umschaut sieht man ... Gesichtern
⊞			670	63	sieht umschaut	60% sich umschaut [...] sieht man dass ...
⊞			672	2	sieht Überall lauter	100% Überall sieht man jetzt darstellende Künstler lauter
⊞			680	8	sieht Überall Gesichter	100% Überall [...] sieht man glückliche|aschfahle und|strahlende bleiche Gesichter die
⊞			681	1	sieht Überall Gesichtern	100% Überall sieht man ... Gesichtern
⊞			994	313	sieht Überall	96% Überall [...] sieht [...] man die
⊞			1023	29	sieht hineinschaut	65% hineinschaut [dann] sieht man dass auch
⊞			1027	4	sieht lauter Gesichter	75% sieht man [...] lauter ... Gesichter
⊞			1028	1	sieht lauter Wohin	100% Wohin ... man sieht ... lauter
⊞			1416	388	sieht lauter	52% sieht [...] man [den Wald|vor] lauter Bäumen
⊞			1418	2	sieht Gesichter Wohin	100% Wohin man [...] sieht fröhliche Gesichter
⊞			1945	527	sieht Gesichter	50% sieht [...] man [...] Gesichter
⊞			2003	58	sieht bloßem	37% sieht man mit bloßem Auge ...
⊞			2166	163	sieht Wohin	63% Wohin [...] man [auch] sieht
⊞			2294	128	sieht gelassener	83% sieht man die|das ... gelassener
⊞			2613	319	sieht Dunkeln	85% Die|die im Dunkeln [...] sieht [...] man nicht
⊞			2633	20	sieht Hebt	75% Hebt man den Blick so sieht man keine
⊞			2762	129	sieht Gesichtern	51% sieht man an|ihren|in|den Gesichtern

Abb. 35 – Ergebnis der KA (Ein Teil des Clusters) – Primärkollokator *sieht*

Die gleiche Analyse des Pronomens *man* wurde auch in der CCDB durchgeführt, in diesem Fall wurden die Synsemantika von der Analyse auch ausgeschlossen.

Abb. 36 – Ergebnis der CCDB – das Pronomen *man*

Die *feine Granularität* in der Einstellung der KA hat verursacht, dass auch etwas unsichere Aussagen über komplexere Kombinationen gezeigt sind (Abb.35). So lassen sich in der Ergebnistabelle mehr als 20 Kookkurrenzen in einem Cluster zum primären Partnerwort *sieht* finden. Dieses Partnerwort befindet sich in der KA auf dem ersten Platz. Im Unterschied dazu lässt sich *sieht* in der CCDB erst auf dem siebten Platz finden und wie es zu sehen ist (Abb. 37), liefert diese Analyse nicht so detailliertes Aufzählen der Kookkurrenzen.

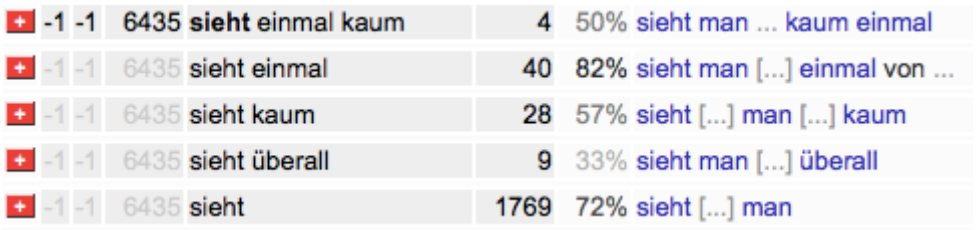


Abb. 37 – Ergebnis der CCDB – Primärkollokator *sieht*

Wenn man die beiden Cluster zu dem Wort *sieht* genauer anschaut, lassen sich auch Unterschiede in den sekundären Partnerwörtern finden. Von drei sekundären

Partnerwörtern in der CCDB (*einmal*, *kaum*, *überall*) lässt sich nur ein gleiches sekundäres Partnerwort auch in der KA finden - und zwar *überall*.

Die Konkordanz, wo nur das primäre Partnerwort *sieht* zu sehen ist, weist nur kleine Differenzen auf (Abb. 38, 39).

Abb. 38 – Ergebnis der KA – Konkordanz – Kollokator *sieht*

Abb. 39 – Ergebnis der CCDB – Konkordanz – Kollokator *sieht*

An dem Kookkurrenzprofil von *man* ist in der CCDB zu erkennen, dass *sieht* und *man* zusammen 1769mal auftreten. 72% dieser Vorkommen folgt dem Muster *sieht […] man*, wohingegen die KA nur 60% angibt. Diese prozentuale Differenz weist wieder auf das sich ständig veränderte Textmaterial der Korpora hin. Die eckigen Klammern zwischen *sieht* und *man* lassen sich in beiden durchgeführten Analysen finden. Sie deuten an, dass die beiden Wörter nicht zwingend unmittelbar nacheinander folgen müssen und dass es sich auch optional ein Ausdruck dazwischen schieben kann.

Für den Vergleich der beiden Analysen wurden auch in diesem Fall die ersten 15 Primärkollokatoren ausgewählt und untereinander verglichen. In der KA handelt es sich um diese Wörter (absteigend geordnet): *sieht, darf, könne, müsse, hört, weiß, wolle, merkt, dann, kennt, müsste, bedenkt, erkennt, denkt* und *Kann*. In der CCDB liefert die Analyse folgende Wörter (absteigend geordnet): *kann, muss, nicht, sollte, muss, hat, sieht, könne, habe, darf, will, könnte, müsse, hört* und *hätte*. Wie es zu sehen ist, befinden sich in der KA unter den ersten 15 Primärkollokatoren nur die Verben, wohingegen die CCDB auch das Negationspartikel *nicht* liefert (Platz 3). In der KA lässt sich *nicht* unter den ersten 100 Primärkollokatoren nicht finden, was ein deutlicher Unterschied zwischen den beiden Analysen ist.

4.5 Analyse eines Numerales

Die nächste Analyse betrifft das Gattungszahlwort ***zweierlei***. Es wurde zuerst die KA durchgeführt, das Korpus liefert für das Numerale *zweierlei* 31 025 Treffer. Für folgende Analyse wurde die *grobe Granularität* ausgewählt, so kann die KA auf Schlagwörter fokussieren, was für den Vergleich der Primärkollokatoren zwischen der KA und CCDB übersichtlichere Ergebnistabelle liefert. Die Funktionswörter wurden von der Analyse ausgeschlossen.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	65383	5773	5773	**Maß**	88% mit zweierlei [...] Maß gemessen
⊞	2	29093	8737	2964	**Hinsicht**	99% in zweierlei [...] Hinsicht
⊞	3	23930	8880	143	**gemessen**	99% mit zweierlei [Mass|Ellen] gemessen wird
⊞	4	8097	8882	2	**messen** warf	100% warf dem Westen vor mit zweierlei ... zu messen
⊞			8972	90	messen	80% nicht mit zweierlei [Mass zu] messen
⊞	5	4199	9785	813	**Gründen**	99% aus zweierlei [...] Gründen
⊞	6	3338	9828	43	**misst**	60% man mit zweierlei Mass|Ellen misst
⊞	7	3264	10597	769	**Weise**	96% auf zweierlei [...] Weise
⊞	8	2780	10785	188	**Mass**	59% mit zweierlei [...] Mass
⊞	9	2577	10793	8	**messe**	87% messe mit zweierlei Ellen
⊞	10	1420	10795	2	**mißt**	50% zweierlei ... mißt
⊞	11	1301	10802	7	**Messen**	100% das Messen mit zweierlei Ellen ...
⊞	12	1125	10967	165	**bedeuten** kann	89% Das kann [...] zweierlei bedeuten
⊞			11038	71	bedeuten	84% würde zweierlei [...] bedeuten
⊞	13	1056	11268	230	**Arten**	96% auf zweierlei [...] Arten von ...
⊞	14	993	11293	25	**Recht** gilt	64% gilt [...] zweierlei Recht
⊞			11606	313	Recht	66% zweierlei [...] Recht
⊞	15	980	11608	2	**hier** Wird	100% Wird hier ... mit zweierlei
⊞			11786	178	hier	76% hier [...] zweierlei

Abb. 40 – Ergebnis der KA – das Numerale *zweierlei*

In der CCDB wurde die gleiche Analyse durchgeführt, die Synsemantika wurden auch ausgeschlossen und die Analyse zielt vor allem auf die ersten 15 Primärkollokatoren.

Abb. 41 – Ergebnis der CCDB – das Numerale *zweierlei*

Unter den ersten 15 Primärkollokatoren in der KA (*Maß, Hinsicht, gemessen, messen, Gründen, misst, Weise, Mass, messe, mißt, Messen, bedeuten, Arten, Recht* und *hier*) findet man mehrmals das Verb *messen* in seinen verschieden Formen (*gemessen, messen, misst, messe, mißt*). Die ersten 15 Primärkollokatoren in der CCDB sind (absteigend geordnet): *Maß, Hinsicht, gemessen, Erstens, messen, Gründen, Weise, erstens, messe, mißt, Recht, Art, Einerseits, Arten* und *beweist*.

Auf dem ersten Platz befindet sich sowohl in der KA als auch in der CCDB als Primärkollokator das Substantiv *Maß*, das in Verbindung mit dem Gattungszahlwort *zweierlei* und mit dem Verb *messen* den Phraseologismus *mit zweierlei Maß messen* bildet. Wenn man aber die Konkordanz in der KA mit der Konkordanz in der CCDB vergleicht, lassen sich wieder Abweichungen finden. Nach der KA folgt 88% der Vorkommen dem Muster *mit zweierlei […] Maß gemessen*, wohingegen die CCDB die Abdeckungsrate 99% angibt.

Die zweit- und dritthäufigsten Primärkollokatoren sind sowohl nach der KA als auch nach der CCDB das Substantiv *Hinsicht* und das Partizip *gemessen*. In diesen Fällen sind auch die syntagmatischen Muster und die Abdeckungsraten gleich.

Die anderen Primärkollokatoren stimmen nur teilweise überein. Z.B. das Adverb *Einerseits* (mit großem Anfangsbuchstaben), das in der CCDB auf dem dreizehnten Platz steht, lässt sich in der KA unter den Primärkollokatoren nicht finden. Mit dem kleinen Anfangsbuchstaben (*einerseits*) befindet sich dieses Adverb in der KA erst auf dem Platz 102, wobei diese Form des Adverbs in der CCDB auf dem Platz 22 steht. Wenn man die konkreten Konkordanzen anschaut, sieht man, dass auch die Abdeckungsrate nicht gleich ist – nach der KA folgt 76% der Vorkommen dem Muster *zweierlei […] einerseits*, wohingegen die CCDB die Abdeckungsrate 100% angibt. Bei diesem Vergleich muss man aber auch auf die Gesamtzahl der Belege, die in dieser Kombination analysiert wurde, Rücksicht nehmen (KA – 13, CCDB – 76).

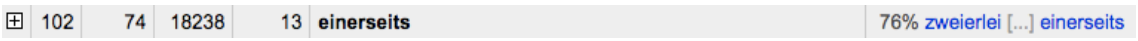


Abb. 42 – Ergebnis der KA – Konkordanz – Kollokator *einerseits*

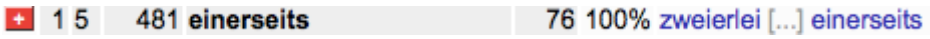


Abb. 43 – Ergebnis der CCDB – Konkordanz – Kollokator *einerseits*

4.6 Analyse eines Verbs

Im dritten Kapitel wurden die Möglichkeiten der KA am Beispiel des Verbs *bestehen* demonstriert (vgl. 3.2). Es wurde bewiesen, dass die KA auch als ein guter Helfer bei der Bestimmung der Valenz dienen kann. In diesem Unterkapitel wird die KA des Verbs ***bestehen*** (*&bestehen*) nochmals durchgeführt und es soll gezeigt werden, ob die KA und die CCDB gleiche Ergebnisse liefern. Es wird auch verglichen, ob die Ergebnisse beider Analysen zur Bestimmung der Valenz von Nutzen sein können. Für diese Zwecke werden die Funktionswörter von beiden Analysen nicht ausgeschlossen.

⊞	1	863680	515115	515115	**aus**	40% besteht [...] aus
⊞	2	567566	515120	5	**darin** Aufgabe Schwierigkeit	40% Schwierigkeit der Aufgabe bestand darin
⊞			515125	5	darin Aufgabe Herausforderung	60% Herausforderung der Aufgabe besteht [...] darin
⊞			515126	1	darin Aufgabe Besonderheit	100% Besonderheit ... Aufgabe bestand darin
⊞			515127	1	darin Aufgabe Vorteil	100% Aufgabe besteht darin ... Vorteil
⊞			515137	10	darin Aufgabe Kunst	30% Aufgabe der p Theaterpädagogik besteht darin die Sprache der Kunst
⊞			521994	6857	darin Aufgabe	44% Die Aufgabe [...] besteht [...] darin die ...
⊞			521995	1	darin Schwierigkeit Herausforderung	100% Schwierigkeit ... Herausforderung ... besteht darin
⊞			521996	1	darin Schwierigkeit Trick	100% Schwierigkeit besteht ... darin ... Trick
⊞			521997	1	darin Schwierigkeit Kunst	100% Schwierigkeit ... Kunst bestehen darin
⊞			523459	1462	darin Schwierigkeit	58% Die Schwierigkeit [...] besteht [...] darin dass die
⊞			523460	1	darin Hauptaufgabe Kunst	100% Hauptaufgabe besteht darin ... Kunst
⊞			524477	1017	darin Hauptaufgabe	41% Die Hauptaufgabe [...] besteht [...] darin die ...
⊞			524481	4	darin Herausforderung Kunst	25% darin besteht ... Herausforderung ... Kunst
⊞			526727	2246	darin Herausforderung	54% Die Herausforderung [...] besteht [...] darin die ...

Abb. 44 – Ergebnis der KA – das Verb *bestehen*

⊞			613247	79219	darin	54% besteht [...] darin dass ...

Abb. 45 – Ergebnis der KA – Konkordanz – Kollokator *darin*

Analysewort: bestehen, Analysetyp 1

⊟	1	1	40683	**aus** Mitgliedern	404	41% Der ... besteht [...] aus [...] Mitgliedern
⊟	1	1	40683	aus Teilen	313	46% besteht [...] aus [zwei|drei] Teilen
⊟	1	1	40683	aus Ortsteilen	95	94% Die Gemeinde besteht [...] aus [den] Ortsteilen
⊟	1	1	40683	aus	20349	40% besteht [...] aus
⊟	-1	4	36511	**darin** Aufgabe Problem	1	100% Aufgabe bestünde ... darin ... Problem
⊟	-1	4	36511	darin Aufgabe	341	51% Die Aufgabe [...] besteht [...] darin
⊟	-1	4	36511	darin Problem	239	66% Das Problem [...] besteht [...] darin daß|dass ...
⊟	-1	4	36511	darin Schwierigkeit	79	53% Die Schwierigkeit [...] besteht [...] darin daß|dass die
⊟	-1	4	36511	darin	5400	54% besteht [...] darin daß|dass ...

Abb. 46 – Ergebnis der CCDB – das Verb *bestehen*

Nach der Untersuchung der KA (Abb. 44, 45) und der CCDB (Abb. 46) ist sichtbar, dass die ersten zwei Primärkollokatoren in den beiden Ergebnistabellen gleich sind (*aus, darin*). Auch das dominante syntagmatische Muster und die Abdeckungsrate für diese zwei Primärkollokatoren stimmen überein – nach beiden Analysen folgt 40% der Vorkommen dem Muster *besteht […] aus* und 54% der Vorkommen dem Muster *besteht […] darin dass/daß*. Im Unterschied zur KA bietet aber die CCDB zu dem Primärkollokator *aus* noch einen Cluster mit Kookkurrenzen höherer Ordnung (*Mitgliedern, Teilen, Ortsteilen*). Die KA gibt nur den Primärkollokator ohne einen Cluster an, obwohl die *feine Granularität* in der Einstellung gewählt wurde. Im Unterschied dazu lässt sich bei dem Primärkollokator *darin* in der KA einen detaillierten Cluster finden, der noch 20 Sekundärkollokatoren enthält (*Aufgabe, Schwierigkeit, Hauptaufgabe, Herausforderung, Besonderheit, Vorteil, Hauptproblem, Trick, Hauptziel, Hauptzweck, Grundidee, Verdienst, Kunst, Pointe, Clou Grundgedanke, Hauptanliegen, Kunstgriff, Crux* und *Tragik*). Die CCDB gibt nur drei Sekundärkollokatoren im Cluster an (*Aufgabe, Problem, Schwierigkeit*).

Bisher wurden nur die ersten zwei Primärkollokatoren in beiden Analysen verglichen, jetzt werden die folgenden 13 Primärkollokatoren beschrieben. In der KA handelt es sich um diese Primärkollokatoren: *Gefahr, Möglichkeit, seit, Handlungsbedarf, feiert, kein, Zweifel, zehnjähriges, 50-jähriges, Verdacht, 25-jähriges, 100-jähriges* und *Einigkeit*. Die CCDB liefert folgende Primärkollokatoren: *Gefahr, Es, Möglichkeit, daß, darauf, kein, Zweifel, dass, seit bereits, Handlungsbedarf, Verdacht* und *Die*. Wie es zu sehen ist, sind die Primärkollokatoren, die durch ein Substantiv vertreten sind, identisch. Die anderen Primärkollokatoren stimmen nicht überein und lassen sich in der gegenüberstellten Analyse auf anderen Plätzen finden. Die Primärkollokatoren *zehnjähriges, 50-jähriges, 25-jähriges* und *100-jähriges*, die in der KA auf dem Platz 10, 11, 13 und 14 stehen, lassen sich in der CCDB sogar gar nicht finden. Dies weist auch darauf hin, dass das sich immer erweiterte Korpus des IDS sehr wahrscheinlich großen Einfluss auf die Ergebnisse der KA hat.

4.7 Analyse eines Adverbs

In diesem Kapitel wurde das Temporaladverb ***damals*** der Analyse unterzogen. In der KA wurden die Funktionswörter von der Analyse ausgeschlossen und die *grobe Granularität*, die auf Schlagwörter fokussiert, wurde ausgewählt. Das Korpus liefert 1 583 694 Treffer und die KA dieser Treffer nimmt etwa 30 Minuten. Erwartet wird eine Liste mit Primärkollokatoren, die wieder mit den Primärkollokatoren der CCDB verglichen werden. Die Analyse in der CCDB wurde auch ohne Synsemantika durchgeführt

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	57125	45927	45927	**heute**	63% damals [wie|und ...] heute
⊞	2	40823	60316	14389	**hieß**	48% hieß [es] damals
⊞	3	35170	76147	15831	**Schon**	99% Schon [...] damals
⊞	4	17169	85546	9399	**alt**	81% war damals [... Jahre] alt
⊞	5	16982	85614	68	**Dr** Prof	48% Prof Dr Helmut Becker 1927-1990 damals an der
⊞			87458	1844	Dr	52% Dr [...] damals
⊞	6	16406	102848	15390	**gewesen**	72% sei damals [...] gewesen
⊞	7	14517	113043	10195	**gesagt**	82% damals [...] gesagt
⊞	8	12149	118347	5304	**galt**	52% galt [...] damals als ...
⊞	9	11988	131310	12963	**ging**	42% ging [es] damals
⊞	10	9083	143144	11834	**hätte**	58% hätte [...] damals
⊞	11	8528	148537	5393	**erinnert**	58% damals [...] erinnert sich
⊞	12	7465	150926	2389	**hiess**	50% hiess [es] damals
⊞	13	7389	154405	3479	**wusste**	52% wusste [...] damals nicht
⊞	14	7310	157449	3044	**dachte**	56% Ich dachte [...] damals
⊞	15	7257	160446	2997	**üblich**	89% wie damals [...] üblich war ...

Abb. 47 – Ergebnis der KA – das Adverb *damals*

Analysewort: damals, Analysetyp **0**

⊞	1	1	24920	**noch** selbständigen	38	100%	der|im damals noch selbständigen Gemeinde ...
⊞	1	1	24920	noch unbekannten	55	96%	dem|den damals [...] noch [...] unbekannten
⊞	1	1	24920	noch existierenden	35	97%	in der damals [...] noch [...] existierenden DDR ...
⊞	1	1	24920	noch	8846	90%	damals [...] noch
⊞	-2	-1	19958	**war** alt	366	70%	war [...] damals [... Jahre] alt und
⊞	-2	-1	19958	war üblich	79	56%	wie es damals [...] üblich [...] war
⊞	-2	-1	19958	war klar	112	40%	Schon|schon damals [...] war [...] klar daß|dass die
⊞	-2	-1	19958	war	11455	67%	war [...] damals
⊞	-1	1	16469	**schon** klar	56	66%	war schon [...] damals [war ...] klar daß
⊞	-1	1	16469	schon gewußt	20	65%	damals schon [...] gewußt daß
⊞	-1	1	16469	schon gewusst	16	56%	damals [...] schon [von ...] gewusst
⊞	-1	1	16469	schon	5899	61%	schon [...] damals
⊞	-1	-1	8587	**Schon** gab	40	100%	Schon damals [...] gab es ...
⊞	-1	-1	8587	Schon galt	20	100%	Schon damals galt ... als
⊞	-1	-1	8587	Schon klar	35	97%	Schon damals war ... klar dass|daß die
⊞	-1	-1	8587	Schon	1318	99%	Schon [...] damals

Abb. 48 – Ergebnis der CCDB – das Adverb *damals*

In diesem Kapitel werden die ersten 15 Primärkollokatoren und ihre syntagmatischen Muster in den beiden Analysen untereinander verglichen. Die ersten 15 Primärkollokatoren in der KA sind (absteigend geordnet): *heute, hieß, Schon, alt, Dr, gewesen, gesagt, galt, ging, hätte, erinnert, hiess, wusste, dachte* und *üblich*. Die Primärkollokatoren in der CCDB sind (absteigend geordnet): *noch, war, schon, Schon, hatte, heute, habe, waren, als, nicht, wurde, hieß, alt, Jahre* und *hatten*. Wie es zu sehen ist, stimmen die Primärkollokatoren fast nicht überein. Die Mehrheit der Primärkollokatoren, die durch ein Verb vertreten sind und die sich unter den ersten 15 Primärkollokatoren in der KA befinden, lassen sich in der CCDB erst auf den unteren Positionen der Ergebnistabelle finden (z.B. *gewesen* – KA #[36]6, CCDB #17; *gesagt* – KA #7, CCDB #19; *galt* – KA #8, CCDB #31; *ging* – KA #9, CCDB #36; *hätte* – KA #10, CCDB #23; *erinnert* – KA #11, CCDB #44; *wusste* – KA #13, CCDB #52; *dachte* – KA #14, CCDB #29).

Wenn man die syntagmatischen Muster der ersten 15 Primärkollokatoren in der KA mit denselben Primärkollokatoren der CCDB vergleicht, sieht man keine Unterschiede. Die Abdeckungsraten heben sich nur in der Größenordnung von höchstens 5% voneinander ab.

[36] # = Nummerierung des Hauptkollokators

4.8 Analyse einer Präposition

Für die Analyse der Präposition wurde die Dativ-Präposition ***angesichts*** ausgewählt. Im Rahmen der KA-Einstellung wurde die *feine Granularität* bewahrt, damit die detaillierte Aufzählung der Kollokatoren in der KA demonstriert werden konnte. Die Funktionswörter wurden für Zwecke der Analyse sowohl in der KA als auch in der CCDB ignoriert.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	68143	13140	13140	**Tatsache**	58% angesichts der Tatsache dass ...
⊞	2	39330	13141	1	**hohen** Arbeitslosigkeit Verschuldung	100% Verschuldung ... angesichts ... hohen Arbeitslosigkeit
⊞			13142	1	hohen Arbeitslosigkeit Staatsverschuldung	100% angesichts ... hohen Arbeitslosigkeit ... Staatsverschuldung
⊞			13143	1	hohen Arbeitslosigkeit Niveaus	100% angesichts ... hohen Niveaus ... Arbeitslosigkeit
⊞			13865	722	hohen Arbeitslosigkeit	60% angesichts der hohen [...] Arbeitslosigkeit
⊞			13968	103	hohen Spritpreise	67% angesichts der hohen Spritpreise
⊞			14162	194	hohen Verschuldung	61% angesichts der hohen [...] Verschuldung der ...
⊞			14283	121	hohen Anteils	50% Angesichts des hohen Anteils von|an ...
⊞			14387	104	hohen Staatsverschuldung	72% angesichts der hohen [...] Staatsverschuldung
⊞			14499	112	hohen Ölpreises	60% angesichts des hohen Ölpreises
⊞			14598	99	hohen Niveaus	64% angesichts des ... hohen [...] Niveaus der
⊞			14644	46	hohen Treibstoffpreise	47% angesichts der hohen Treibstoffpreise
⊞			14669	25	hohen Kursniveaus	64% angesichts des hohen Kursniveaus
⊞			14696	27	hohen Ausländeranteils	62% angesichts des hohen Ausländeranteils in ...
⊞			14727	31	hohen Preisniveaus	64% angesichts des hohen [...] Preisniveaus
⊞			14775	48	hohen Jugendarbeitslosigkeit	58% angesichts der hohen Jugendarbeitslosigkeit in ...

Abb. 49 – Ergebnis der KA – die Präposition *angesichts*

⊞			24477	9521	hohen	60% angesichts [der|des] hohen

Abb. 50 – Ergebnis der KA – Konkordanz – Kollokator *hohen*

Analysewort: angesichts, Analysetyp **0**

⊞	2	2	11298	**Tatsache** erscheint	13	100%	erscheint [...] angesichts der Tatsache daß die
⊞	2	2	11298	Tatsache Wunder	12	100%	kein|Kein Wunder [...] angesichts der Tatsache dass|daß die
⊞	2	2	11298	Tatsache allem	39	97%	vor|Vor allem [...] angesichts der Tatsache dass|daß die
⊞	2	2	11298	Tatsache	1337	96%	angesichts der Tatsache dass|daß ...
⊞	2	2	8732	**hohen** Arbeitslosigkeit	147	97%	angesichts der hohen [...] Arbeitslosigkeit
⊞	2	2	8732	hohen Alters	37	97%	angesichts ihres|seines|des hohen Alters der ...
⊞	2	2	8732	hohen Ölpreise	23	100%	angesichts der hohen Ölpreise
⊞	2	2	8732	hohen	1538	96%	angesichts [der] hohen
⊞	2	4	8053	**Lage** schwierigen	137	99%	angesichts der schwierigen [...] Lage
⊞	2	4	8053	Lage prekären	67	97%	angesichts der prekären [...] Lage der ...
⊞	2	4	8053	Lage unsicheren	42	100%	sich angesichts der unsicheren [...] Lage
⊞	2	4	8053	Lage	1773	92%	angesichts der ... Lage
⊞	1	1	7003	**solcher** Zahlen Wunder	1	100%	Wunder ... angesichts solcher Zahlen
⊞	1	1	7003	solcher Zahlen	50	100%	angesichts solcher [...] Zahlen
⊞	1	1	7003	solcher Wunder	16	100%	Kein Wunder [daß|dass ...] angesichts solcher
⊞	1	1	7003	solcher Aussichten	13	100%	sich angesichts solcher Aussichten
⊞	1	1	7003	solcher	931	93%	angesichts [...] solcher

Abb. 51 – Ergebnis der CCDB – die Präposition *angesichts*

Wie es in der Ergebnistabelle zu sehen ist, liefern die beiden Analysen auf dem ersten Platz den gleichen Primärkollokator – *Tatsache.* Im Unterschied zu der KA gibt aber die CCDB zu diesem Primärkollokator auch sekundäre Kollokatoren (*erscheint, Wunder, allem*) an. Wenn man die beiden Konkordanzen mit dem syntagmatischen Muster anschaut (Abb. 52, 53), sieht man, dass die Abdeckungsrate nicht gleich ist – nach der KA folgt 58% der Vorkommen dem Muster *angesichts der Tatsache, dass*, wohingegen die CCDB die Abdeckungsrate 96% angibt.

Abb. 52 – Ergebnis der KA – Konkordanz – Kollokator *Tatsache*

Abb. 53 – Ergebnis der CCDB – Konkordanz – Kollokator *Tatsache*

Der Primärkollokator *hohen* ist in beiden durchgeführten Analysen auf dem zweiten Platz, aber die Menge der Sekundärkollokatoren stimmt nicht überein. Die CCDB liefert nur 3 Sekundärkollokatoren, dahingegen liefert die KA 20 Sekundärkollokatoren in einem Cluster. Auch die Abdeckungsraten unterscheiden sich voneinander (Abb. 54, 55) - nach der KA folgt 60% der Vorkommen dem Muster *angesichts [der/des] hohen*, wohingegen die CCDB die Abdeckungsrate 96% bei dem Muster *angesichts [der] hohen* angibt.

Abb. 54 – Ergebnis der KA – Konkordanz – Kollokator *hohen*

Abb. 55 – Ergebnis der CCDB – Konkordanz – Kollokator *hohen*

Zu den Differenzen zwischen den Abdeckungsraten kommt es auch im Fall des dritten Primärkollokators *Lage* (Abb. 56, 57). Nach der KA folgt 60% der Vorkommen dem Muster *angesichts [der…] Lage*, wohingegen die CCDB die Abdeckungsrate 92% bei dem Muster *angesichts der… Lage* angibt. Genauso wie im obigen Beispiel lässt sich auch bei dem Primärkollokator *Lage* in der KA einen detaillierten Cluster mit 20

Sekundärkollokatoren finden, in der CCDB lassen sich wieder nur drei Sekundärkollokatoren finden.

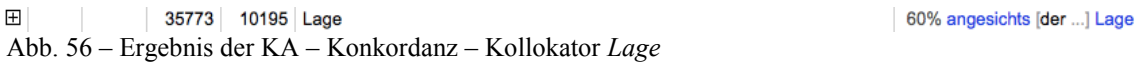


Abb. 56 – Ergebnis der KA – Konkordanz – Kollokator *Lage*

Abb. 57 – Ergebnis der CCDB – Konkordanz – Kollokator *Lage*

Von dem vierten Platz haben die Primärkollokatoren in der KA und in der CCDB nicht die gleiche Reihenfolge. Die nächsten 12 Primärkollokatoren in der KA sind (absteigend geordnet): *steigender, Haushaltslage, drohenden, solcher, angespannten, Situation, wachsenden, leerer, steigenden, Kassen, bevorstehenden* und *schwierigen*. In der CCDB handelt es sich um diese Wörter: *solcher, bevorstehenden, drohenden, leerer, Haushaltslage, sei, angespannten, Situation, Kassen, wachsenden, steigender* und *Finanzlage*. Wie es zu sehen ist, sind die Primärkollokatoren, die durch ein Substantiv vertreten sind, identisch. Das Substantiv *Finanzlage*, das unter den ersten 15 Primärkollokatoren der KA nicht zu sehen ist, befindet sich in der Ergebnistabelle auf dem Platz 17. Die anderen Primärkollokatoren stimmen auch fast überein, nur das Adjektiv *schwierig* ist in der Reihe nicht zu sehen, weil es erst auf Platz 16 steht. Allgemein lässt sich also zu dieser Analyse sagen, dass die Reihenfolge und die Abdeckungsraten nicht völlig übereinstimmen, aber die wichtigsten Primärkollokatoren in beiden Analysen ähnlich sind.

4.9 Analyse einer Konjunktion

Die folgende Analyse betrifft die konzessive Konjunktion ***wenngleich***. In der KA wurde zuerst die Grundeinstellung (*Granularität: fein*) bewahrt. Die Analyse hat in diesem Fall eine riesige Menge von Tertiärkollokatoren geliefert, die sich in der CCDB gar nicht finden lassen. Für einen übersichtlicheren Vergleich der Ergebnisse wurde die Einstellung der Granularität in der KA geändert (*Granularität: mittel*). Die Funktionswörter wurden in diesem Fall von der Analyse ausgeschlossen. Das Korpus liefert für die Konjunktion *wenngleich* 70 678 Treffer.

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	4941	8867	8867	**auch**	51% wenngleich [...] auch
⊞	2	2695	8868	1	**noch** immer Details unklar	100% wenngleich noch immer ... Details unklar
⊞			8870	2	noch immer Details	50% wenngleich noch immer ... Details
⊞			8872	2	noch immer ausbaufähig	50% wenngleich immer noch ausbaufähig
⊞			8873	1	noch immer unklar	100% wenngleich ... immer noch unklar
⊞			9489	616	noch immer	34% wenngleich [...] immer noch
⊞			9497	8	noch Selbsterfahrungsexperiments scherzte	100% Beginn des Selbsterfahrungsexperiments klar wenngleich ich da noch scherzte mich
⊞			9507	10	noch aussteht	70% wenngleich der|die ... noch aussteht
⊞			9508	1	noch Details unklar	100% wenngleich noch ... Details unklar
⊞			9530	22	noch Details	50% wenngleich [einige] Details noch nicht ...
⊞			9539	9	noch feststeht	66% nicht ... wenngleich [...] noch nicht|nichts Konkretes feststeht ob
⊞			9544	5	noch ausbaufähig	60% wenngleich natürlich ... noch ausbaufähig
⊞			9555	11	noch Einzelheiten	45% wenngleich [er] Einzelheiten [...] noch
⊞			9568	13	noch unklar	46% Wenngleich [...] noch [...] unklar ist ...
⊞			9570	2	noch hinreichend	50% Wenngleich noch ... hinreichend
⊞			13148	3578	noch	55% wenngleich [...] noch

Abb. 58 – Ergebnis der KA – die Konjunktion *wenngleich*

Die Funktionswörter werden auch nicht in die CCDB-Analyse einbezogen.

Analysewort: wenngleich, Analysetyp 0

⊞	1	4	3045	**nicht** unbedingt sonderlich	1	100%	wenngleich nicht unbedingt sonderlich
⊞	1	4	3045	nicht unbedingt	56	78%	wenngleich [...] nicht [...] unbedingt
⊞	1	4	3045	nicht sonderlich	24	75%	wenngleich [...] nicht [...] sonderlich
⊞	1	4	3045	nicht gravierend	8	50%	wenngleich [...] nicht so gravierend
⊞	1	4	3045	nicht	3523	65%	wenngleich [...] nicht
⊞	1	2	1793	**auch** hier indirekt	1	100%	auch hier ... wenngleich indirekt
⊞	1	2	1793	auch hier	132	62%	wenngleich [...] auch [...] hier
⊞	1	2	1793	auch indirekt	8	62%	auch [...] wenngleich [...] indirekt
⊞	1	2	1793	auch seltener	7	57%	auch [...] wenngleich seltener
⊞	1	2	1793	auch	2394	57%	wenngleich [...] auch
⊞	1	5	876	**noch** immer bislang	1	100%	immer ... wenngleich ... bislang noch
⊞	1	5	876	noch immer	224	34%	wenngleich [...] immer noch
⊞	1	5	876	noch bislang	13	53%	wenngleich [...] bislang noch
⊞	1	5	876	noch Details	7	71%	wenngleich ... Details [...] noch nicht ...
⊞	1	5	876	noch	1380	62%	wenngleich [...] noch

Abb. 59 – Ergebnis der CCDB – die Konjunktion *wenngleich*

Die *mittlere Granularität* in der Einstellung der KA hat verursacht, dass auch etwas unsichere Aussagen über komplexere Kombinationen gezeigt werden (Abb.58). So lassen sich in der Ergebnistabelle mehrere Kookkurrenzen in einem Cluster z.B. zum primären Partnerwort *noch* finden. Wenn man die beiden Cluster (Abb. 58, 59) zu dem Wort *noch* genauer anschaut, lassen sich auch Unterschiede in den sekundären Partnerwörtern finden. Im Unterschied zu den drei sekundären Partnerwörtern in der CCDB (*immer, bislang, Details*) lassen sich in der KA noch weitere finden (*Selbsterfahrungsexperiment, aussteht, feststeht, ausbaufähig, Einzelheiten, unklar* und *hinreichend*). Der Sekundärkollokator *bislang* erscheint aber in der KA, im Unterschied zu der CCDB, nicht.

Auf dem ersten Platz in der KA befindet sich der Primärkollokator *auch*, zu dem die Analyse keinen Cluster mit Kookkurrenzen liefert. Im Unterschied dazu liefert die CCDB zu diesem Partnerwort einen Cluster mit Kookkurrenzen höherer Ordnung (genau drei Sekundärkollokatoren – *hier, indirekt, seltener*).

Auf dem ersten Platz in der CCDB sieht man den Primärkollokator *nicht* (mit dem syntagmatischen Muster *wenngleich […] nicht* mit der Abdeckungsrate 65%). In der KA findet man diesen Kollokator unter den Primärkollokatoren gar nicht, was ein deutlicher Unterschied zwischen beiden Analysen ist.

Für einen übersichtlichen Vergleich der beiden Analysen wurden auch in diesem Fall die ersten 15 Primärkollokatoren ausgewählt und untereinander verglichen. In der KA handelt es sich um diese Wörter (absteigend geordnet): *auch, noch, natürlich, durchaus, immer, eher, einräumt, sehr, zufrieden, zugibt, wohl, hier, Niveau, ganz* und *recht*. In der CCDB liefert die Analyse folgende Wörter (absteigend geordnet): *nicht, auch, noch, durchaus, zufrieden, immer, natürlich, hier, ist, sehr, eher, ganz, einräumt, gut* und *unterschiedlichem*. Die Mehrheit der ersten Primärkollokatoren stimmt zwar überein, aber ihre Reihenfolge ist nicht identisch. Z.B. das einzige Substantiv *Niveau*, das unter den ersten 15 Primärkollokatoren der KA auf dem Platz 13 zu finden ist, lässt sich in der CCDB erst auf dem Platz 20 finden und die Abdeckungsrate des syntagmatischen Musters *wenngleich auf ... Niveau* ist in der CCDB um 10% höher als in der KA (also 82%).

4.10 Analyse einer Interjektion

Obwohl die Interjektionen ein typisches Merkmal der gesprochenen Sprache sind, lassen sie sich auch in Korpora der geschriebenen Sprache finden. In der schriftlichen Kommunikation dienen sie vor allem als Stilmittel der Mündlichkeit. Für folgende Analyse wurde die Interjektion ***ach*** ausgewählt. Es werden wieder die ersten 15 Primärkollokatoren in den beiden Analysen untereinander verglichen und es wird auf die Differenzen gewiesen.

In der Einstellung der KA wurde zuerst die *ganz grobe Granularität* ausgewählt und die Funktionswörter wurden von der Analyse ausgeschlossen. Bei solcher Einstellung liefert die KA folgende Ergebnisliste:

	#	LLR	kumul.	Häufig	Kookkurrenzen	syntagmatische Muster
⊞	1	134893	24613	24613	**ja**	85% Ach [...] ja
⊞	2	15330	27270	2657	**Gott**	80% Ach [du lieber] Gott
⊞	3	14302	29240	1970	**nein**	83% Ach [...] nein
⊞	4	12056	30194	954	**nee**	83% Ach [...] nee
⊞	5	5616	32724	2530	**Herr**	79% - Ach [...] Herr
⊞	6	5086	34262	1538	**schön**	84% Ach [wie das] schön
⊞	7	4604	34848	586	**komm**	85% Ach [...] komm
⊞	8	4487	35125	277	**Gottchen**	87% Ach [...] Gottchen
⊞	9	4151	35811	686	**lieber**	82% Ach [du] lieber
⊞	10	4076	36678	867	**liebe**	90% Ach [du] liebe Zeit
⊞	11	2830	37312	634	**hab**	86% Ach [ich] hab ich|in meinem
⊞	12	2744	37751	439	**Quatsch**	92% Ach [...] Quatsch
⊞	13	2727	38054	303	**Güte**	90% Ach du meine Güte
⊞	14	2614	38459	405	**Seelen**	90% zwei|Zwei Seelen [wohnen] ach in meiner
⊞	15	2397	38550	91	**Schnucki**	95% Schnucki ach Schnucki

Abb. 60 – Ergebnis der KA

Die ersten 15 Primärkollokatoren sind (absteigend geordnet): *ja, Gott, nein, nee, Herr, schön, komm, Gottchen, lieber, liebe, hab, Quatsch, Güte, Seelen* und *Schnucki*.

In der CCDB wurden die Synsemantika von der Analyse auch ausgeschlossen. Die ersten 15 Primärkollokatoren in der CCDB sind (absteigend geordnet): *ja, Seelen, nein, Gott, Brust, Habe, nee, wohnen, Schnucki, Zwei, schön, sagen, lieber, bösen* und *Gottchen*.

Analysewort: ach, Analysetyp 0

	1	1	5212	**ja** ist war	3	33% ist ... war ... ach ja
	1	1	5212	ja ist noch	3	33% ist ja noch ... ach
	1	1	5212	ja ist	117	30% ach [das] ist [...] ja
	1	1	5212	ja war noch	11	63% war [das|doch] noch [was] ach ja
	1	1	5212	ja war	47	40% da war [doch noch was] ach [...] ja
	1	1	5212	ja noch	61	45% war das noch [...] ach [...] ja
	1	1	5212	ja	927	85% ach [...] ja
	-2	-1	1696	**Seelen** Brust wohnen Zwei	38	89% Zwei Seelen wohnen ach in seiner|meiner Brust
	-2	-1	1696	Seelen Brust wohnen	55	87% Zwei Seelen wohnen ach in meiner|seiner Brust
	-2	-1	1696	Seelen Brust Zwei	50	100% Zwei Seelen [wohnen] ach in seiner|meiner Brust
	-2	-1	1696	Seelen Brust	103	93% Zwei|zwei Seelen [wohnen] ach in meiner|seiner Brust
	-2	-1	1696	Seelen wohnen Zwei	56	91% Zwei Seelen wohnen ach in meiner
	-2	-1	1696	Seelen wohnen	81	83% Zwei Seelen wohnen ach in meiner
	-2	-1	1696	Seelen Zwei	78	100% Zwei Seelen [wohnen] ach in meiner
	-2	-1	1696	Seelen	158	93% zwei|Zwei Seelen [wohnen] ach in meiner
	1	1	1507	**nein** nicht antwortete wohl	1	100% antwortete ... wohl ach nein ... nicht
	1	1	1507	nein nicht wohl	3	33% wohl ach nein ... nicht
	1	1	1507	nein nicht	27	44% ach nein [... ich ...] nicht
	1	1	1507	nein antwortete	5	80% antwortete [...] ach [...] nein
	1	1	1507	nein wohl	8	50% ach nein ... ich wohl
	1	1	1507	nein	189	82% ach [...] nein

Abb. 61 – Ergebnis der CCDB

Wie es zu sehen ist, haben die Analysen andere Reihenfolge der Primärkollokatoren geliefert, einige Primärkollokatoren sind sogar ganz unterschiedlich. Z.B. die Primärkollokatoren, die durch ein Substantiv vertreten sind und die sich unter den ersten 15 Primärkollokatoren in der KA befinden, lassen sich in der CCDB erst auf den unteren Positionen der Ergebnistabelle finden (z.B. *Herr* – KA #[37]5, CCDB #24; *Quatsch* – KA #12, CCDB #27; *Güte* – KA #13, CCDB #99). Der Primärkollokator *komm*, der in der KA auf dem siebten Platz steht, lässt sich in der CCDB erst auf der Position 109 finden.

Die gleiche KA wurde nochmals durchgeführt, diesmal wurde aber die *feine Granularität* ausgewählt. Diese Analyse soll auf den Vergleich der Cluster der KA und der CCDB zielen. Für diese Zwecke wurde das Substantiv *Seelen* ausgewählt, das sich in der KA auf dem Platz 14 befindet (Abb. 62), dahingegen steht *Seelen* in der CCDB auf dem zweiten Platz (Abb. 64).

Die hohen Prozentangaben bei den syntagmatischen Mustern in beiden Ergebnistabellen weisen darauf hin, dass der Primärkollokator *Seelen* zusammen mit der Interjektion *ach* feste Wortverbindung bildet. In diesem Fall handelt es sich um die

[37] # = Nummerierung des Hauptkollokators

Wortverbindung *zwei Seelen wohnen ach in (meiner/seiner) Brust*, die von Goethes Faust stammt.

⊞	14	2614	38057	3	**Seelen** Brust wohnen Zwei Goethe	66% Goethe Zwei Seelen wohnen ach ... in meiner Brust
⊞			38059	2	Seelen Brust wohnen Zwei Goethes	100% Goethes [...] Zwei Seelen wohnen ach in meiner Brust
⊞			38170	111	Seelen Brust wohnen Zwei	95% Zwei [...] Seelen wohnen [...] ach in meiner Brust
⊞			38200	30	Seelen Brust wohnen	60% - zwei Seelen [...] wohnen [...] ach in meiner|seiner Brust
⊞			38201	1	Seelen Brust Zwei Faustens wohnten	100% Zwei Seelen wohnten ach ... Faustens Brust
⊞			38205	4	Seelen Brust Zwei wohnten	100% Zwei Seelen wohnten ach in seiner Brust
⊞			38236	31	Seelen Brust Zwei	100% Zwei Seelen [schlagen] ach in meiner|seiner Brust
⊞			38237	1	Seelen Brust Faustens	100% Seelen ... ach ... Faustens Brust
⊞			38239	2	Seelen Brust Goethes Faust	100% Goethes Faust zwei Seelen ach in ... Brust
⊞			38245	6	Seelen Brust wohnten	33% Seelen wohnten ach in seiner Brust
⊞			38246	1	Seelen Brust Faust	100% Faust ... Seelen ach ... Brust
⊞			38293	47	Seelen Brust	87% zwei Seelen [...] ach in meiner|seiner Brust

Abb. 62 – Ergebnis der KA

⊞			38294	1	Seelen wohnen Zwei Goethe	100% Goethe Zwei Seelen wohnen ach
⊞			38295	1	Seelen wohnen Zwei Goethes Faust	100% Goethes Faust Zwei Seelen wohnen ach
⊞			38296	1	Seelen wohnen Zwei Goethes	100% Goethes Zwei Seelen wohnen ach
⊞			38377	81	Seelen wohnen Zwei	97% Zwei Seelen wohnen [...] ach
⊞			38378	1	Seelen wohnen Goethe	100% Goethe ... Seelen ach ... wohnen
⊞			38379	1	Seelen wohnen Faustens	100% wohnen ach ... Seelen ... Faustens
⊞			38398	19	Seelen wohnen	63% zwei Seelen wohnen ach in
⊞			38402	4	Seelen Zwei wohnten	100% Zwei Seelen wohnten [...] ach
⊞			38404	2	Seelen Zwei Faust	50% Zwei Seelen ach ... Faust
⊞			38420	16	Seelen Zwei	100% Zwei Seelen [...] ach
⊞			38421	1	Seelen Goethe	100% Goethe ... Seelen ach
⊞			38422	1	Seelen Goethes Faust	100% Goethes Faust ... Seelen ach
⊞			38423	1	Seelen wohnten	100% Seelen wohnten ach
⊞			38424	1	Seelen Faust	100% Faust ... Seelen ach
⊞			38459	35	Seelen	51% zwei Seelen [die] ach

Abb. 63 – Ergebnis der KA

⊞	-2	-1	1696	**Seelen** Brust wohnen Zwei	38	89% Zwei Seelen wohnen ach in seiner|meiner Brust
⊞	-2	-1	1696	Seelen Brust wohnen	55	87% Zwei Seelen wohnen ach in meiner|seiner Brust
⊞	-2	-1	1696	Seelen Brust Zwei	50	100% Zwei Seelen [wohnen] ach in seiner|meiner Brust
⊞	-2	-1	1696	Seelen Brust	103	93% Zwei|zwei Seelen [wohnen] ach in meiner|seiner Brust
⊞	-2	-1	1696	Seelen wohnen Zwei	56	91% Zwei Seelen wohnen ach in meiner
⊞	-2	-1	1696	Seelen wohnen	81	83% Zwei Seelen wohnen ach in meiner
⊞	-2	-1	1696	Seelen Zwei	78	100% Zwei Seelen [wohnen] ach in meiner
⊞	-2	-1	1696	Seelen	158	93% zwei|Zwei Seelen [wohnen] ach in meiner

Abb. 64 – Ergebnis der CCDB

Wie es zu sehen ist, liefert die KA dank der feinen Granularität präzisere Aufzählung der weiteren Kollokatoren. Allgemein lässt sich also sagen, dass für detailliertere Untersuchung der Verwendung eines Phraseologismus die KA geeigneter

zu sein scheint. Der Nutzer bekommt in der KA mehrere Beispiele, aus denen er die präziseren Hypothesen von der Verwendung der Phraseologismen ableiten kann.

5 Schlusswort

Das Thema der Korpuslinguistik findet die Verfasserin als Germanistikstudentin sehr wichtig. In heutiger Zeit stellen die Textkorpora eine großartige Möglichkeit dar, wie man die Sprache in ihrer Entwicklung beobachten und analysieren kann. Ohne Textkorpora hätte man kaum einen so breiten Zugriff zu der authentischen Datenmenge, die vom Gebrauch natürlicher Sprachen ausgeht. Dank der Korpora können die Linguisten den Schriftsprachgebrauch dokumentieren, die gewonnenen Daten auf wissenschaftstheoretischer Ebene reflektieren und in die Diskussion der linguistischen Theoriebildung einbringen.

Die vorliegende Arbeit setzte sich zum Ziel die Differenzen und Gemeinsamkeiten der CCDB und der KA zu erörtern. Die CCDB ist eine statische Sammlung von Kookkurrenzanalyseergebnissen, die auf Grund der KA herausgebildet wurde. Dahingegen ist die KA nicht statisch und ihre Ergebnisse ändern sich genauso, wie sich das DeReKo ändert und kontinuierlich erweitert (über 6 Milliarden Textwörter - Stand: Dezember 2013; über 31 Milliarden Textwörter - Stand: März 2017). Die Tatsache, dass sich das DeReKo kontinuierlich erweitert und dass immer neue Texte zum Bestandteil der Korpussammlung werden, hat auch einen Einfluss auf die Ergebnisse, die die KA liefert. Aus diesem Grund waren die Primärkollokatoren in den durchgeführten Analysen nicht immer identisch und die Reihenfolge der Kollokatoren war in vielen Fällen unterschiedlich. Die Größe und das sich ständig veränderte Textmaterial der Korpora können die Ergebnisse der KA mit größter Wahrscheinlichkeit beeinflussen, was im Rahmen dieser Arbeit überprüft wurde.

Die CCDB ist eine statische Sammlung, daraus folgt, dass sie auch keine weitere Entwicklung der Sprache widerspiegeln kann und die Ausdrücke, die sie an ersten Positionen angibt, müssen in der Zukunft nicht so frequentiert und aktuell sein. Im Unterschied dazu sind die Ergebnisse der KA aktuell und reflektieren die Sprachentwicklung (z.B. Verwendung der Neologismen und Anglizismen, Ausdrücke der Jugendsprache). Die KA arbeitet mit größerer Textmenge, was ihre Ergebnisse von den CCDB-Ergebnissen unterscheidet.

Der nächste wichtige Unterschied zwischen der KA und der CCDB betrifft die Parametereinstellung. Die KA ist auf beliebige COSMAS-Suchobjekte anwendbar mit optionaler Lemmatisierung, variabler Kontextgröße, ggf. automatischer Fokussierung auf den Kontext mit dem stärksten Kohäsionswert, variabler Zuverlässigkeit (d.h. Signifikanz des ersten Kookkurrenzpartners), variabler Granularität (d.h. Signifikanz der Kookkurrenzpartner, die für die Ermittlung von Mehrworteinheiten berücksichtigt werden), variabler Zuordnung von Belegen bei Mehrworteinheiten und Berechnung von syntagmatischen Mustern zu jedem Kookkurrenzcluster[38] (vgl. 3.1). Dahingegen darf man in der CCDB die Einstellungen nicht ändern. Das hängt damit zusammen, dass die einzelnen KA-n in der CCDB schon gespeichert sind und man ruft sie nur auf. Der Nutzer darf in der CCDB nur wählen, ob auch Synsemantika ein Bestandteil der Analyse sein werden oder nicht.

Der nächste Unterschied hängt mit der Auswahl der Quellen zusammen. Am Anfang der KA hat der Nutzer die Möglichkeit, das Archiv auszuwählen (es gibt insgesamt 18 Archive der COSMAS II – Korpora). Diese Möglichkeit hat der Nutzer im Rahmen der CCDB nicht, da sich die Analysen in der CCDB nicht mehr beeinflussen lassen.

Die nächste, rein praktische Differenz betrifft den Zeitaufwand. Da jede KA des IDS neu durchgeführt werden muss, kann sie auch mehr als eine Stunde dauern. Das betrifft vor allem die KA der hochfrequentierten Wörter (vgl. 4.3 – Analyse des Artikels *der*). Wegen des technischen Aufwands kann die anspruchsvolle KA auch scheitern, was zur Wiederholung der Operation führt. Im Unterschied dazu sind alle Analysen in der CCDB gespeichert, der Nutzer ruft sie nur ab und die Ergebnistabelle ist gleich zu sehen. Was den Zeitaufwand betrifft, scheint die CCDB benutzerfreundlicher zu sein.

Die Ergebnisse des praktischen Teils der vorliegenden Arbeit weisen darauf hin, dass die KA und die CCDB in vieler Hinsicht Differenzen aufweisen und dass sie nicht als übereinstimmende Analysen betrachtet werden können.

[38] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 27.6.2017] .

6 Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Korpuslinguistik und beschreibt näher zwei Applikationen, die die Daten vom Korpus DeReKo bearbeiten und die sich auf den Corpus-Driven-Zugang stützen. Es handelt sich um die Kookkurrenzanalyse (KA) und die Kookkurrenzdatenbank (CCDB).

Die CCDB kann als eine Datenbank charakterisiert werden, die auf der Basis von DeReKo angelegt ist, und in der die Ergebnisse von Analysen zu mehr als 220.000 Grundformen gespeichert sind. Im Unterschied zu der KA, deren Basiskorpus kontinuierlich erweitert wird, enthält die CCDB Analysen, die schon berechnet wurden und deren Parametereinstellung der Nutzer nicht mehr beeinflussen kann. Im Rahmen der vorliegenden Arbeit wurden die Spezifika der KA und der CCDB anschaulich an ausgewählten Beispielen demonstriert und detailliert beschrieben (vgl. 3.1, 3.3), wobei auch die Termini aus dem Bereich der KA (z.B. *Granularität*, *Autofokus*, *Lemmatisierung*) definiert wurden (vgl. 3.2). Die Ergebnisse, die man im Rahmen der KA oder der CCDB gewinnt, weisen darauf hin, dass sie die linguistischen Hypothesen entweder bestätigen oder widerlegen können, oder dass sie auch zum Gewinn neuer Theorien nützlich sind. Wie in dieser Arbeit auch gewiesen wurde, ermöglichen die beiden korpusanalytischen Methoden einen empirischen Zugang zu Massendaten, indem sie Präferenzsetzungen vornehmen und hochfrequente Belegmengen ordnen und strukturieren.

Der empirische Teil der Arbeit setzte sich zum Ziel, die Ergebnisse der KA und der CCDB zu vergleichen und die Differenzen und Gemeinsamkeiten zu erörtern. Zu diesem Zweck dienten flektierbare und nicht flektierbare Wortarten (das Substantiv *Gefahr*, das Adjektiv *schwarz*, der Artikel *der*, das Pronomen *man*, das Numerale *zweierlei*, das Verb *bestehen*, das Adverb *damals*, die Präposition *angesichts*, die Konjunktion *wenngleich* und die Interjektion *ach*), die sowohl in der KA als auch in der CCDB analysiert wurden. Für jedes ausgewählte Wort wurden die Konkordanzen erstellt, Wortliste mit Frequenzangaben erzeugt und Kookkurrenzen berechnet. Um auf die Gemeinsamkeiten und Differenzen zwischen der CCDB und der KA aufmerksam zu machen, wurden diese Angaben in jedem Unterkapitel untereinander verglichen. Die vorliegende Arbeit hat bestätigt, dass beide Analysen in vielen Fällen ganz

unterschiedliche Kollokatoren liefern. Es wurde auch bewiesen, dass beide Analysenmethoden oft unterschiedliche Reihenfolge oder verschiedene Wortformen von Kollokatoren angeben.

Beide Analysenmethoden ermöglichen u. a. eine empirische Erfassung usueller Wortverbindungen als Kandidaten für Mehrworteinheiten der deutschen Gegenwartssprache (Phraseologismen, Redewendungen, Sprichwörter, kommunikative Formeln, Funktionsverbgefüge usw.).[39] Diese Behauptung wurde in der Diplomarbeit bestätigt. In der Regel treten als Ergebnis der KA und der Analyse in der CCDB viele Phraseologismen auf. Diese Tatsache hängt sehr wahrscheinlich damit zusammen, dass die Phraseologismen Stabilität aufweisen. Die Identifikation von Phraseologismen wurde in der Arbeit anschaulich gezeigt und an mehreren Beispielen (z.B. *schwa*rz vgl.4.2) bestätigt. Wie es in der Arbeit bewiesen wurde, liefert die CCDB die Kollokatoren nur im kleineren Maß als die KA. Der Nutzer bekommt im Rahmen der KA detaillierte Aufzählung der Kollokatoren, von denen er die präziseren Hypothesen über die formale Seite der Phraseologismen ableiten kann. Da die KA (im Unterschied zu der CCDB) dem Nutzer ermöglicht, die Parametereinstellung zu ändern (z.B. *Granularität: fein, mittel, grob*), scheint sie für die nähere Untersuchung der gefundenen Phraseologismen geeigneter zu sein (vgl. 4.10).

Die vorliegende Arbeit hat auch bestätigt, dass die KA und die CCDB für die Bestimmung der Valenz dienen können. Die Valenzbestimmung wurde am Beispiel des Verbs *bestehen* gezeigt (vgl. 4.6).

Der empirische Teil der Arbeit weist darauf hin, dass der Korpusumfang, auf dem die beiden Analysenmethoden basieren (KA – DeReKo – über 31 Milliarden Textwörter - Stand: März 2017; CCDB – DeReKo – 2,2 Milliarden Textwörter), die Ergebnisse der Analysen verändern kann. Aus diesem Grund können die KA und die CCDB nicht als dieselbe Analysenmethoden betrachtet werden.

[39] Die Informationen wurden aus http://www1.ids-mannheim.de/kl/projekte/methoden/ka.html übernommen [abgerufen am 29.5.2017].

7 Shrnutí

Předkládaná práce se zabývá korpusovou lingvistikou a blíže popisuje dvě aplikace, které zpracovávají data z korpusu DeReKo přístupem corpus-driven. Jedná se o kookurenční analýzu (KA) a Kookurenční databanku (CCDB).

CCDB je charakterizována jako databanka, která vznikla na bázi DeReKo a ve které jsou uloženy analýzy k více než 220.000 základním slovním tvarům. Na rozdíl od KA, jejíž výsledky se zakládají na současném jazykovém korpusu DeReKo, jehož rozsah se kontinuálně rozšiřuje, obsahuje CCDB analýzy, které již byly provedeny a jejichž nastavení uživatel nemůže dále ovlivnit.

V rámci předkládané práce bylo poukázáno na specifika KA a CCDB (viz 3.1, 3.3), přičemž byly definovány pojmy z oblasti KA (např. *granularita, autofokus, lematizace*). Výsledky, které uživatel získá v rámci KA a CCDB, poukazují na to, že obě aplikace mohou být užitečné jak pro potvrzení či vyvrácení lingvistických hypotéz, tak k vytvoření nových teorií o jazykových jevech.

Empirická část diplomové práce si kladla za cíl porovnat výsledky KA a CCDB, přičemž mělo být poukázáno na rozdíly a podobnosti mezi oběma aplikacemi. K tomuto účelu bylo vybráno vždy jedno slovo od každého slovního druhu (podstatné jméno *Gefahr*, přídavné jméno *schwarz*, gramatický člen *der*, zájmeno *man*, číslovka *zweierlei*, sloveso *bestehen*, příslovce *damals*, předložka *angesichts*, spojka *wenngleich* a citoslovce *ach*). Tato slova byla následně analyzována v CCDB a byla provedena jejich KA. Pro každé vybrané slovo byly vytvořeny a zobrazeny konkordance, kookurence a údaje o čestnosti a frekvenci nalezených kolokátorů. Aby mohlo být poukázáno na rozdíly a podobnosti obou analýz, byly mezi sebou tyto údaje v každé podkapitole porovnány a díky obrazovým přílohám doloženy. Předkládaná práce prokázala, že se v mnoha případech v obou analýzách objevují zcela odlišné kolokátory. Stejně tak bylo dokázáno, že obě aplikace často udávají rozdílné pořadí nebo rozdílné slovní tvary nalezených kolokátorů.

Obě aplikace mají kromě jiného umožňovat evidenci víceslovných slovních spojení německého jazyka (frazeologismů, rčení, přísloví atd.). Toto tvrzení bylo v rámci diplomové práce potvrzeno. Zpravidla se ve výsledcích KA a CCDB vyskytuje vícero frazeologismů. Tento jev souvisí velmi pravděpodobně s tím, že jsou frazeologismy v jazyce pevně ukotveny (vykazují stabilitu). To, že je KA a CCDB vhodná pro identifikace frazeologismů, bylo v předkládané práci ukázáno na vícero příkladech (např. *schwarz* viz 4.2). CCDB poskytuje uživateli menší počet kolokátorů než KA, z tohoto důvodu považuje autorka této práce KA jako vhodnější aplikaci k detailnímu prozkoumání užití frazeologismů. Díky variabilnímu nastavení KA (např. možnost nastavení granularity) se uživateli naskýtá možnost detailnějšího prozkoumání frazeologismů, díky kterému je možné vytvářet preciznější hypotézy o jejich formální stránce (viz 4.10).

KA a CCDB mají kromě jiného sloužit také k určení či potvrzení valence, toto tvrzení bylo v rámci práce také potvrzeno a na příkladu *bestehen* demonstrováno (viz 4.6).

Empirická část této práce poukazuje na to, že rozsah korpusu, ze kterého obě aplikace vychází (KA – DeReKo – přes 31 Miliard textových slov – stav: březen 2017; CCDB – DeReKo – 2,2 Miliardy textových slov), může měnit výsledky analýz. Výsledky potvrzují, že data získaná pomocí CCDB a KA jsou v mnoha ohledech odlišná (např. rozdílné kolokátory, rozdílné pořadí kolokátorů, rozdílné slovní tvary nalezených kolokátorů) a potvrzují tak autorčinu hypotézu, že rozdílná velikost korpusu hraje ve výsledcích obou aplikací zásadní roli. Z tohoto důvodu nemohou být KA a CCDB považovány za totožné aplikace.

8 Literatur und Quellen

Literatur:

BELICA, Cyril und Kathrin STEYER. *Korpusanalytische Zugänge zu sprachlichem Usus.* In: VACHKOVÁ, Marie (Hg.). Beiträge zur bilingualen Lexikographie. Praha: Univerzita Karlova v Praze, Filozofická fakulta, 2008, S. 7-24. ISBN 9788073082178.

ČERMÁK, František. *Korpusová lingvistika Praha 2011*. Vydání 1. Praha: Nakladatelství Lidové noviny, 2011, 323 s. ISBN 978-80-7422-115-6.

ČERMÁK, František a Renata BLATNÁ, ed. *Korpusová lingvistika - stav a modelové přístupy*. Praha: NLN, Nakladatelství Lidové noviny, 2006. Studie z korpusové lingvistiky. ISBN 80-7106-861-6.

ČERMÁK, František, Jana KLÍMOVÁ a Vladimír PETKEVIČ. *Studie z korpusové lingvistiky*. 1. vyd. Praha: Karolinum, 2000, 531 s. ISBN 80-7184-893-x.

ENGELBERG, Stefan und Lothar LEMNITZER. *Einführung in die Lexikographie und Wörterbuchbenutzung.* Tübingen: Stauffenburg, 2001.

KUNZE, Claudia und Lothar LEMNITZER. *Computerlexikographie. Eine Einführung.* Tübingen: Narr [E-Book], 2007.

LEMNITZER, Lothar a Heike ZINSMEISTER. *Korpuslinguistik eine Einführung*. Tübingen: Narr, 2006. ISBN 3823362100.

LEMNITZER, Lothar und Heike ZINSMEISTER. *Korpuslinguistik: Eine Einführung.* 3., überarbeitete und erweiterte Auflage. Tübingen: Narr Francke Attempto, 2015. ISBN 978-3-8233-6886-1.

PERKUHN, Rainer, Holger KEIBEL und Marc KUPIETZ. *Korpuslinguistik.* Paderborn: Wilhelm Fink, 2012. ISBN 978-3-8252-3433-1.

SCHERER, Carmen. *Korpuslinguistik.* Winter, Heidelberg 2006, ISBN 3-8253-5164-5.

SCHERER, Carmen. *Korpuslinguistik*. 2., aktualisierte Auflage. Heidelberg, Neckar: Universitätsverlag Winter GmbH Heidelberg, 2014. ISBN 9783825363147.

SCHWITALLA, Johannes. a Werner. WEGSTEIN. *Korpuslinguistik.* Tübingen: Max Niemeyer Verlag, 2005. ISBN 3-484-73064-1.

ŠULC, Michal. *Korpusová lingvistika: první vstup*. Praha: Karolinum, 1999. ISBN 80-7184-847-6.

Internetquellen:

BELICA, Cyril: *Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In:* Korpusinstrumente in Lehre und Forschung / Corpora: strumenti per la didattica e la ricerca / Corpus Tools in Teaching and Research. *Bozen : alpha beta piccadilly Verlag, im Druck.* Stand 29.5.2017, abgerufen unter:

http://corpora.ids-mannheim.de/SemProx.pdf

BERMAN, Stephen. *Allgemeine Begriffe der Korpuslinguistik*, Stand 29.5.2017, abgerufen unter:

http://homepage.rub.de/Stephen.Berman/Korpuslinguistik/Allgemeines.html

BICKEL, Balthasar, Mathias, JENNY und Steven, MORAN. *Was ist Allgemeine Sprachwissenschaft?*, Stand 29.5.2017, abgerufen unter:

http://www.comparativelinguistics.uzh.ch/dam/jcr:53751112-97b0-4792-9d0a-39e53741effd/whatisgenerallinguistics20150901_de.pdf

Blogs Lehre Dresden Center for Digital Linguistics, Stand 17.1.2017, abgerufen unter:

http://linguistik.zih.tu-dresden.de/lehre/blogs/blog/2015/02/04/kookkurrenzanalyse-ein-blog-ueber-die-definition-und-funktion-einer-korpuslinguistischen-methode/

http://linguistik.zih.tu-dresden.de/lehre/blogs/blog/2015/02/23/zusammenfassung-von-semantische-naehe-als-aehnlichkeit-von-kookkurrenzprofilen/

BOPP, Sebastian. *Einführung in die Korpuslinguistik mit DeReKo und COSMAS II*. 2., aktualisierte und korrigierte Fassung, Stand 29.5.2017, abgerufen unter:

https://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/mitarbeiter/stelspass/materialien_lehrveranstaltungen/korpuslinguistik_dereko_cosmas2_bopp.pdf

IDS - Institut für deutsche Sprache, Stand 17.7.2017, abgerufen unter:

https://cosmas2.ids-mannheim.de/cosmas2-web/

http://corpora.ids-mannheim.de/ccdb/

http://www1.ids-mannheim.de/kl.html

http://www1.ids-mannheim.de/kl/projekte/korpora/

http://www1.ids-mannheim.de/direktion/kl/projekte/korpora/archiv.html

http://www.ids-mannheim.de/cosmas2/projekt/referenz/archive.html

http://www1.ids-mannheim.de/kl/misc/tutorial.html

http://corpora.ids-mannheim.de/SemProx.pdf

http://www.ids-mannheim.de/cosmas2/projekt/einsteiger/was.html

http://www.ids-mannheim.de/cosmas2/projekt/hilfe/quick.html

http://www1.ids-mannheim.de/fileadmin/ids/Downloads/flyer-ccdb.pdf

http://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_LingMeth/Skript_06.pdf

Webseite – Noah Bubenhofer, Stand 17.7.2017, abgerufen unter:

https://www.bubenhofer.com/korpuslinguistik/kurs/index.php?id=cosmas_client_kookk.html

Wikipedia, Stand 29.5.2017, abgerufen unter:

https://de.wikipedia.org/wiki/Korpuslinguistik

https://de.wikipedia.org/wiki/Textkorpus#cite_note-1