– Ergänzungen zu –

Korpuslinguistik

Rainer Perkuhn / Holger Keibel / Marc Kupietz (2012):
Korpuslinguistik. Paderborn: Fink.
(Reihe LIBAC – Linguistik für Bachelor 3433).

Stand: 18. Juni 2012

Inhalt

E2 Elementare Korpuseinheiten und einfache Recherche

E2.9 Recherche- und Analysesystem des IDS: Einfache Recherche mit COSMAS II

Um die in diesem Abschnitt vorgestellte praktische Übung nacharbeiten zu können, müssen Sie über einen Computer, einen Internetzugang und -Browser verfügen.

Wenn Sie die in Kapitel 2 unseres Buches vorgestellten Konzepte praktisch anwenden wollen, können Sie beliebige Recherchesysteme ausprobieren. Für die Recherche in den Korpora des IDS ist allerdings erforderlich, dass Sie sich für COSMAS II, das Recherchesystem des IDS, registrieren – da der in dieser Form eingeschränkte Zugang zu den Daten in den Lizenzverträgen mit den Textgebern so vereinbart wurde. Gehen Sie dazu auf die Website »https://cosmas2.ids-mannheim.de/cosmas2-web/«. Lesen Sie die Nutzungsbedingungen aufmerksam durch. Die Registrierung ist kostenlos. Sie müssen jedoch bestätigen, dass Sie verstanden haben, dass die Nutzung der Daten nur für (sprach-)wissenschaftliche, insbesondere nicht-kommerzielle Zwecke erlaubt ist. Vor diesem Hintergrund muss der Zugang zu den Daten überwacht werden. Versuche, an dieser Vereinbarung vorbei duch systematische Suchanfragen komplette Dokumente herunterzuladen, werden als Missbrauch eingestuft und können notfalls sogar gerichtliche Schritte nach sich ziehen.

Wir werden in diesem Abschnitt nicht das komplette Recherchesystem COSMAS II und dessen Suchanfragesyntax erklären können. Dazu sei auf die Online-Hilfe verwiesen. Ziel ist es nur, Sie mit der elementaren Bedienung vertraut zu machen und Sie zu animieren, vieles selber auszuprobieren. Dazu wollen wir zunächst eine minimale Sitzung durchspielen. Danach werden wir einige Besonderheiten aufgreifen, die wir bereits vorher skizziert hatten oder auf die wir später zurückgreifen wollen.

E2.9.1 Eine minimale COSMAS-Sitzung

Starten Sie eine COSMAS II-Sitzung. So wie Sie über die obere orange, horizontale Navigationsleiste zur Anmeldung geleitet werden, können Sie auch eine Recherchesitzung beginnen. In der linken vertikalen Menuleiste werden Ihnen die jeweiligen Schritte anboten, die grundsätzlich oder als nächstes möglich sind. Innerhalb einer Recherche sind dies zunächst »Archiv« und »Korpus«. Für Ihre spätere Forschungsarbeit sollten Sie sich auf jeden Fall mit diesen Konzepten, insbesondere mit dem des benutzerdefinierten, virtuellen Korpus vertraut machen. Der Einfachheit halber wählen wir jetzt hier »W – Archiv der geschriebenen Sprache« als Archiv und »W-gesamt – alle Korpora des Archivs W (mit Neuakquisitionen)« als bereits vordefiniertes Korpus. Während einer Rechercheanfrage begleitet Sie nun ein grau hinterlegter Bereich, in dem festgehalten wird,

Archiv, Korpus

Auswahl über Doppelklick

welche Auswahl Sie getroffen haben bzw. welche Eingaben Sie gemacht haben. Sie sehen hier z. B. das ausgewählte Archiv, das ausgewählte Korpus und ein Feld für die Suchanfrage. In diesem grauen Bereich können Sie aber nichts unmittelbar eingeben oder ändern. Dies geschieht immer in entsprechenden Bereichen, zu denen Sie notfalls immer wieder über die linke Navigationsliste gelangen können. So auch bei der Formulierung der Suchanfrage. Für dessen Eingabe wird Ihnen unterhalb des grauen Bereichs ein großes Feld angeboten. Schreiben nun dort eine Zeichenkette hinein, nach der Sie suchen wollen. Verändern Sie sonst nichts und starten Sie dann den Suchvorgang. In Abhängigkeit von Ihren persönlichen Einstellungen kann jetzt ein Zwischenschritt erfolgen, der Aufbau und die Anzeige einer Wortformliste, deren Erklärung wir jetzt aber zunächst zurückstellen. Wählen Sie in dem Fall einfach »Ergebnisse«. Ob mit oder ohne Wortliste wird Ihnen dann als nächstes das Suchergebnis angezeigt, zunächst in seiner sehr groben Form, die davon abhängt, welche Ansicht bei Ihnen eingestellt ist. In der linken Navigationsleiste können Sie andere Ansichten ausprobieren. Diese Ansichten sind jeweils definiert über Entstehungsbedingungen der Texte, sei es Zeit, Region oder Quelle.

Suchanfrage

Wortformenliste
Ergebnisse

Im grauen Protokollfenster wird unten rechts die Trefferanzahl vermerkt. Unter diesem Fenster werden Ihnen nun vier Möglichkeiten angeboten, wie Sie weiter mit der Ergebnismenge verfahren können. Die ersten beiden Punkte sind selbsterklärend: Über »Gesamt-KWIC« bzw. »Gesamt-Volltext« gelanden Sie zur Anzeige von Konkordanzen bzw. größeren Textabschnitten, sowie wie wir sie im Buch beschrieben haben.

Trefferanzahl

KWIC-/
Volltextanzeige

Als kleine Aufgabe zwischendurch können Sie jetzt versuchen, Inhaltswörter als Wortformen (keine Eigennamen) suchen zu lassen, von denen Sie erwarten, dass Sie eine hohe Trefferzahl verbuchen können! Welche Wortform hat bei Ihnen die meisten Treffer?

Um nun aber einmal eine COSMAS II-Minimalsitzung, sozusagen »in a nutshell« bis zum Ende durchgelaufen zu sein, wollen wir noch die Exportfunktion ausprobieren: Wählen Sie den dritten der unter dem grauen Kästchen angebotenen Punkte, »Export«, aus. Im nun erscheinenden Dialog sollten Sie einen eigenen Dateinamen vergeben, die Auswahl auf die ersten 100 Treffer einschränken und das Häkchen bei Volltext herausnehmen – die letzten beide Punkte ohne jeglichen methodischen Anspruch, nur, damit es schneller geht. Wenn Sie dann über »exportieren« den eigentlichen Vorgang anstoßen, wird auf Ihrem Rechner lokal eine Datei angelegt, die das Suchergebnis in der von Ihnen ausgewählten Form enthält. Damit hat man ein für viele Zwecke recht brauchbares Textdokument, mit dem weiterarbeiten kann. Aber wie kann man jetzt eigentlich weiter arbeiten? Vieles an strukturierter Information ist nun nur noch aus dem Layout herauszulesen, für eine systematische Auswertung aber verlorengegangen.

Exportieren

E2.9.2 Wortformenlisten, reguläre Ausdrücke, Lemmatisierung

Wortformenliste

Lassen Sie uns aber nun die Erklärung zu Wortformenlisten nachholen, die wir gerade zurückgestellt haben. Um alle wieder am selben Punkt einsteigen zu lassen, starten wir eine neue Suchanfrage und hängen ein »*« an den Suchausdruck. Sie ahnen sicher schon, dass jetzt reguläre Ausdrücke ins Spiel kommen. Wenn ein Suchausdruck quasi eine Variation enthält, wird zunächst eine Wortformenliste erzeugt, die die Wortformen enthält, die auf das Suchmuster passen und die in den Korpora belegt sind.

reguläre Ausdrücke

Eingeschränkte primitive reguläre Ausdrücke in COSMAS II:	
*	beliebig viele beliebige Zeichen
?	genau ein beliebiges Zeichen
+	ein beliebiges oder kein Zeichen

In dem Zwischenschritt wird dem Anfragenden dann die Möglichkeit angeboten, sich diese Liste anzuschauen und bei Bedarf einzugreifen: Formen, die nicht interessieren, können deaktiviert werden; fehlende Formen müssen anderweitig ergänzt werden. Die Variation kann auch implizit durch zugelassende Varianten der Groß- und Kleinschreibung oder bei den diakritischen Zeichen verursacht worden sein – und kann somit für entsprechende Wortformenlisten (auch bei Ihrer ersten Suche nach Wortformen ohne reguläre Ausdrücke) verantwortlich (gewesen) sein. Dies hängt allerdings von den persönlichen Einstellungen ab, die Sie unter Optionen/Suche überprüfen und verändern können. Da direkt unter der Rubrik auch Optionen zur »Lemmatisierung« angeboten werden, ahnen Sie sicher auch, dass Sie in COSMAS II einen entsprechenden Operator einsetzen können. Mit Hilfe eines vorangestelltem »&« können Sie auf die einer Grundform zugeordneten Wortformen zurückgreifen. Diese Zuordnung hat ein automatischer Lemmatisierer erstellt, der keinen Kontext berücksichtigt und auch nicht darauf ausgelegt ist, Mehrdeutigkeiten aufzulösen. Desweiteren ist er so tolerant ausgelegt, dass er möglichst gut mit den authentischen Daten zurecht kommt, was allerdings zur Folge haben kann, dass er in manchen Zweifelsfällen daneben liegt. Eine gute Strategie ist es deshalb hier auch, den Operator einzusetzen, sich die Wortformenliste, manchmal auch Expansionsliste genannt, anzuschauen und zu prüfen, ob die aufgelisteten Wortformen dem entsprechen, was man erwartet hat.

Suchoptionen
Lemmatisierung
Expansionsliste

Wie weit der Lemmatisierer »Lemmatisieren« deutet, hängt von der Option Lemmatisierung ab. Dort ist einstellbar, ob nur Flexion, oder auch eine oder mehrere der Wortbildungen Kompostion, andere Bildungsformen oder Spezialfälle (z. B. Bindestrichschreibweisen) mit unter Lemmatisierung gefasst werden sollen.

E2.9.3 Sortieren

Gehen wir noch einmal zurück zu einem Suchergebnis mit einer hohen Trefferzahl. Wenn Sie Glück (oder Pech?) haben, stehen Ihnen nun hun-

derte oder sogar tausende von KWIC-Zeilen oder Volltexte zur weiteren Auswertung zur Verfügung. Aber wie sollen Sie dieser Masse Herr werden? Die verschiedenen Ergebnisansichten können nur bei einer ersten und sehr oberflächlichen Einschätzung helfen, die Masse der Texte erschlägt den Anfragenden eher als dass sie ihn weiter bringt. Sind dann nicht doch kleinere Korpora besser als große? Dann hätte man doch nur kleine Treffermengen, die man gut handhaben kann! Auf den ersten Blick mag man dem zwar beipflichten. Wenn man aber länger darüber nachdenkt, wiegen die Gefahren eines kleines Korpus deutlich schwerer. Die Qualität der Stichprobe kann nicht unbedingt sichergestellt werden, viele interessante, aber seltene Phänomene werden dann einfach nicht erfasst. Diese Treffer gibt es dann sozusagen nicht (vgl. Bemerkungen zu Typ-II-Fehler in unserem Buch). Das Problem der großen Treffermengen lässt sich aber vielleicht doch in den Griff kriegen – mit allen Vorteilen einer größeren zugrundegelegten Stichprobe.

Um große Treffermenge besser handhaben zu können, wäre es schön, diese nach irgendwelche Kriterien sortieren zu können. Das hieße, sie in eine Reihenfolge zu bringen, durch die man schneller an bestimmte Phänomene gelangen kann, durch die bestimmte Phänomene leichter zugänglich sind bzw. deutlicher zum Vorschein kommen.

Was soll man aber nun nach welchen Kriterien sortieren? Wenn bereits der gesuchte Begriff Varianz enthält – etwa über Alternative, Lemmatisierung oder reguläre Elemente –, dann ist auch sinnvoll, über das Trefferobjekt zu sortieren. Daneben wäre aber sicher auch interessant, die Wörter in der Umgebung für eine Sortierung heranzuziehen, im einfachsten Fall jeweils die Wörter auf einer bestimmten Position davor oder danach. Sortieren könnte man nach Anzahl der Zeichen oder alphabetisch-lexikographisch. Hierbei begegnen uns wieder alle Probleme der Tokenisierung und der Gleichheit von Zeichen (Groß-/Kleinschreibung, Umlaute, diakritische Zeichen, ß, fremde Zeichen aus nicht-lateinischen Alphabeten usw.). Eine besonders interessante und aussichtsreiche Variante besteht in der Sortierung nicht alphabetisch »normal« sondern die Wörter rückläufig zu betrachten, d. h. dass die Zeichen vom Ende eines Wortes her genommen für die Einsortierung Ausschlag gebend sind. Dies ist sehr ergiebig, um Wortbildungsphänomene gebündelt betrachten zu können, z. B. alle Derivationssuffixe oder die gleichen Zweitglieder bei Kompositareihenbildung jeweils zusammen. Standardmäßig – und so auch in COSMAS II – wird aber nur vorwärts zeichengenau sortiert.

Alphabetische Sortierung

Diese sehr interessante Möglichkeit verbirgt sich unter der Option KWIC/Volltext: Stellen Sie – nach einer Suche z. B. einer Körperteilbezeichnung – unter Alphabetische Sortierung das erste Kriterium auf 1. Vorgänger (»Übernehmen«) und lassen Sie die Gesamt KWIC alphabetisch sortieren (bei Gesamt-KWIC in der linken Navigationsleiste: »alphabetisch«). Was stellen Sie fest, wenn Sie die KWIC-Zeilen der Reihe nach überblicken? Zumindest ansatzweise sieht man jetzt Regelmäßigkeiten, viele erscheinen zwar trivial (die Artikel und andere Funktionswörter),

aber manche auch interessant (»gebrochenes« bei »Nasenbein«), je nachdem, welches Erkenntnisinteresse den Forschenden angetrieben hat. So schön es auch scheinen mag, ein wenig Struktur zu erahnen, grundsätzlich bleiben zwei Bedenken. Egal, welchen Mediums wir uns bedienen, können wir trotzdem nur einen kleinen Ausschnitt überschauen. Wie aber das im sichtbaren Ausschnitt Beobachtete zum Gesamtverhalten der beteiligten Wörter in Beziehung zu setzen ist, bleibt unklar. Vielleicht gibt es ja ganz viele Wörter, die sich genauso verhalten? Und was ist mit Wörtern, die sowieso oft im Gesamtkorpus vorkommen? Ist es bei diesen – wie etwa bei den Funktionswörtern – nicht sogar zu erwarten, dass sich eine Abfolge oft wiederholt? Bräuchten wir dann nicht eine relative Einschätzung, wie viele Vorkommen in der Nähe des Suchworts zu verzeichnen sind im Vergleich zu für wie viele dies nicht zutrifft? Genau darauf zielen die letzten Kapitel in unserem Buch.

E2.9.4 Komplexere Suchanfragen

Häufig stehen nicht nur einzelne Wörter im Mittelpunkt des Interesses, sondern ganze Mehrwortfolgen. In einem ersten Schritt wollen wir deshalb jetzt versuchen, nach Zweiwortfolgen zu suchen. Denken Sie sich eine einfache Adjektiv-Substantiv Folge aus (etwa »gebrochenes Nasenbein«, »rote Ampel«, »grüne Minna«, »blondes Haar«, »runder Tisch«, ...) und formulieren Sie eine Suchanfrage, die genau nach diesen beiden Wörtern in der Folge nacheinander sucht. Falls Sie die beiden Wörter mit Leerzeichen dazwischen in Anführungszeichen gesetzt haben, haben Sie sich vielleicht davon irritieren lassen, wie es in der Klammer angegeben war – oder sich zu sehr an *google* orientiert. Die Suchanfragesprache von COSMAS II mag zwar zu Beginn etwas ungewohnt und an manchen Stellen umständlich erscheinen, ist aber in vielen Punkten wesentlich mächtiger für die Suche in den IDS Korpora und speziell auf linguistische Fragestellungen zugeschnitten. Die Suchanfragesyntax muss aber ein wenig erlernt werden, sie ist aber auch nicht weiter schwer.

Abstandsoperator

Wenn die Positionen der Wörter zueinander in der Anfrage spezifiziert werden sollen, benötigen wir einen Abstandsoperator. Im einfachsten Fall kann man den Wortabstand nach rechts von »1« durch die Option unter dem Eingabefeld (weggelassener Operator bedeutet /+w1) einschalten, Anführungszeichen braucht man nicht.

Als Beispiel für eine Anfrage mit Wortabstand können Sie die von Ihnen oben gewählte entsprechend dem folgenden Beispiel anpassen und ausprobieren:

```
großer /+w2:5 Tisch
```

Allgemein hat der Abstandsoperator die schematische Form

```
/<richtung><einheit><minimum>:<maximum>
↑     ↑         ↑         ↑     ↑     ↑
/     +         w         2     :     5
```

mit den Bestandteilen:

einschließender Abstand

/	bedeutet »einschließender Abstand«
<richtung>	kann »+« oder »–« sein
»+«	bedeutet Abstand nach rechts
»-«	Abstand nach links
weglassen	bedeutet Abstand nach rechts oder links (quasi ein Fenster der doppelten Länge um das Wort herum)
<einheit>	ist »w« für Wort, »s« für Satz, »p« für Absatz
<minimum>	und
<maximum>	sind selbsterklärend, der Doppelpunkt gehört zwingend zum minimum, beides zusammen kann weggelassen werden mit der Bedeutung »nur Angabe des maximum«: bis höchstens so viele Wörter Abstand

Bei der Beispielanfrage wird nach der Wortform »großer« gesucht, nach der mindestens zwei, höchstens aber fünf Positionen später die Wortform »Tisch« folgt. Probieren Sie verschiedene Anfragen mit verschiedenen Abständen aus. Was fällt Ihnen insbesondere bei großen Zahlen für den Maximalabstand auf? Wenn Sie die Spanne zwischen zwei Wörtern auf Positionen gezählt in Wörtern beziehen, haben Sie damit nicht ausgeschlossen, dass die Spanne Satzgrenzen überschreitet. Wenn man diese Treffer vermeiden will, kann man sich des Satzabstandes bedienen. Die Anfrage

großer /+s0 Tisch

bedeutet, dass die Wortform »großer« gesucht wird, auf die irgendwo danach im selben Satz die Wortform »Tisch« folgt – eine weitere Einschränkung auf Wortabstände müsste hierbei dann zusätzlich angegeben werden.

Was bedeutet wohl die Anfrage »großer /s0 Tisch«? – Die Wortform »großer«, mit der zusammen irgendwo im selben Satz die Wortform »Tisch« vorkommt. Als Abstand kann also auch die »0« sinnvoll eingesetzt werden, in manchen Fällen sogar auch bezogen auf Wortabstände. Denken Sie mal darüber nach, für welche Art von Anfragen dies einsetzbar ist!

Nachdem wir die erste Art des Abstands »einschließender Abstand« genannt haben, erwarten Sie sicher zu Recht, dass es auch eine andere Form des Abstandes gibt. Mit einem kleinen einfachen Beispiel wollen wir die beiden Formen gegenüberstellen. Die Suche nach der Wortform »Kohl« führt zu vielen Treffern des Gemüses, aber auch des Bundeskanzlers. Möchte man (näherungsweise) nur die letzten, hat man die Möglichkeit »Bundeskanzler /+w1:1 Kohl« suchen zu lassen. Man kann sich die Anfrage als von links nach rechts zu interpretierende Relativkonstruktion vorstellen: »Suche die Wortform »Bundeskanzler«, bei der ein Wort später die Wortform »Kohl« steht«. Die gleiche Menge von Textstellen würde

man erhalten, wenn die Anfrage andersherum aufgebaut wäre: »Kohl /-w1:1 Bundeskanzler« – »Suche Wortform »Kohl«, bei ein Wort davor die Wortform »Bundeskanzler« steht«. Was ist aber, wenn man aber gerade diese Treffer nicht haben möchte, zwar die Wortform »Kohl«, aber eben nicht mit der Wortform »Bundeskanzler« davor. Um diese ausschließen (!) zu können, gibt es den naheliegenden Operator für den »ausschließenden Abstand«: »%«. Mit diesem sieht dann die entsprechende Anfrage folgendermaßen aus: »Kohl %-w1:1 Bundeskanzler«. Dadurch werden eben genau die Fundstellen aus der Treffermenge ausgeschlossen, bei denen der zweite Bestandteil sich innerhalb des Abstands zum ersten Bestandteil befindet. Umschrieben hieße diese Anfrage: »Suche die Wortform »Kohl«, bei ein Wort davor nicht die Wortform »Bundeskanzler« steht«. Um das Beispiel inhaltlich abzuschließen, müsste genau genommen natürlich eine alternative Aufzählung von Wörtern davor ausgeschlossen werden (etwa »Bundeskanzler oder Kanzler oder Helmut oder Regierung«) – das überlassen wir Ihnen aber zum Ausprobieren.

ausschließenden Abstand

Um aber die Bedeutung dieses Operators noch einmal zu durchdenken: Wonach würde denn mit »Bundeskanzler %+w1:1 Kohl« gesucht werden? Im obigen Fall des einschließenden Abstands war die Reihenfolge der Bestandteile ja unerheblich. Und hier?

Probieren Sie es aus und erklären, was Sie beobachten!

Gefunden wird hierbei die Wortform »Bundeskanzler«, hinter der nicht die Wortform »Kohl« steht, also Verwendungen ohne Namen oder mit Namen von anderen Bundeskanzlern!

Abschließend wollen wir Ihnen noch eine kleine Übersicht über weitere Operatoren zusammenfassen.

weitere Operatoren

Operator	Kommentar
$	unmittelbar vor Wort, um auf jeden Fall Groß- und Kleinschreibung bei diesem Wort zuzulassen
()	um Teilausdrücke zu klammern, Operatoren immer zwischen zwei Ausdrücken, a op b op c nicht ohne Zusatzregeln auflösbar, deshalb explizite Klammerung (a op b) op c, um erst a und b miteinander zu verknüpfen, oder a op (b op c), um erst b und c miteinander zu verknüpfen
UND ODER	um Teilausdrücke auf Texte bezogen zu verknüpfen, z. B. »Maus UND Labor« um Texte zu finden, in denen beide Wörter vorkommen, »anfangen ODER beginnen« um Texte zu finden, in denen das eine oder andere Wort vorkommt
NICHT	um Texte zu finden, die zwar auf den vorhergehenden, aber nicht auf den folgenden Suchausdruck passen
	die logischen Operatoren ebenfalls klein geschrieben und englische Formen möglich; um alle diese als Wort selbst zu suchen, sind sie in Anführungszeichen zu setzen

Mit diesem kleinen Überblick über Recherche und der kurzen Einführung in COSMAS II sollten Sie für die ersten Sitzungen gerüstet sein. Wählen Sie sich Ihre Aufgabenstellung und versuchen Sie, dies mit dem Ihnen hier zur Verfügung gestellten Instrumentarium zu bearbeiten.

E2.10 Aufgaben zum praktischen Teil

1. Suchen Sie in den Korpora des IDS Wortformen, bei denen Sie jeweils eine hohe Trefferzahl erwarten, und zwar
   a) zwei unmittelbar aufeinander folgende Wortformen, von denen beide Inhaltswort, aber kein Eigenname sind,
   b) zwei Wortformen innerhalb eines Satzes, von denen beide Inhaltswort, aber kein Eigenname sind.

   Welches Wortpaar hat bei Ihnen die meisten Treffer?
2. Wählen Sie ein Wort (z. B. ein Farbadjektiv) und recherchieren Sie in den Korpora des IDS mit den auf der Begleitseite vorgestellten Möglichkeiten. Schreiben Sie einen kurzen Artikel zu dem Wort, etwa in Form eines Wörterbucheintrags, die Struktur ist von Ihrer Seite frei wählbar. Heben Sie besonders hervor, wenn Sie Verwendungen des Wortes gefunden haben, die Sie selbst oder auch von einem herkömmlichen Wörterbuch nicht erwartet hätten.
   Vergleichen Sie Ihren Artikel mit dem Ergebnis einer *Google*-Suchanfrage.