EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät

Deutsches Seminar

Regina Zieleke

Handreichung zur Korpusrecherche

im Deutschen Referenzkorpus (DeReKo) über COSMAS II$_{web}$

Das Deutsche Referenzkorpus (DeReKo) gehört zu den Korpora der geschriebenen Gegenwartssprache des Mannheimer Instituts für Deutsche Sprache. Diese bilden mit 50,6 Milliarden Wörtern (Stand 02.02.2021) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit.[1]

Für die elektronische Recherche sind die Korpora über die Web-Applikation COSMAS II zugänglich. Die Verwendung der Korpora erfordert eine (kostenfreie) **Registrierung**, bei der Sie angeben, dass Sie die Daten zu Forschungszwecken verwenden möchten.

1. Grundlagen: Korpus – wie und warum? 2
2. Die Wahl des Korpus 4
3. Die Suchanfrage in COSMAS II 6
3.1. Allgemeine Anfragesyntax in COSMAS II 6
3.2. Morphosyntaktische Tags 10
4. Die Ergebnisse 12
4.1. Export der Ergebnisse 12
4.2. Annotation der Ergebnisse 14
5. Die Auswertung 15
5.1. Deskriptive Auswertung 15
5.2. Darstellung der Korpusuntersuchung in der wissenschaftlichen Arbeit 17
6. Nützliche Links und weiterführende Literatur 18

[1] https://www1.ids-mannheim.de/kl/projekte/korpora.html

1. Grundlagen: Korpus – wie und warum?

Ein Korpus (Neutrum, Plural: Korpora) ist eine endliche, aber erweiterbare Sammlung bereits bestehender Sprachdaten wie z.B. Zeitungstexten, die als Grundlage für verschiedene linguistische Studien dienen kann. Korpora sind in der Regel digitalisiert und maschinell verfügbar, repräsentativ zusammengestellt sowie mit Metadaten (z.B. Entstehungsdatum, linguistische Informationen) versetzt.

Mithilfe von Korpusdaten können **sprachliche Phänomene in natürlichen Kontexten** der Sprachproduktion untersucht werden, und zwar strukturiert und unabhängig von der Intuition der/s Forscherin/s. Das bringt zwei Vorteile für linguistische Untersuchungen mit sich: Zum einen bietet die ‚Natürlichkeit' der Sprachdaten die Möglichkeit, das untersuchte Phänomen mithilfe einer **qualitativen Analyse** umfassender zu beschreiben, als es durch reine Introspektion möglich wäre. Zum anderen erlaubt die hohe Anzahl der Sprachdaten eine **quantitative Analyse** sprachlicher Eigenschaften und Tendenzen. So kann herausgefunden werden, ob introspektiv ermittelte Eigenschaften überhaupt im tatsächlichen Sprachgebrauch verwendet werden und, wenn ja, wie oft.

An dieser Stelle sei jedoch auf eine wichtige **Einschränkung** von Korpusdaten hingewiesen: Bloß, weil ein bestimmtes Phänomen den Korpusdaten zufolge nicht (oder nicht häufig) produziert wird, bedeutet das nicht, dass es nicht produziert werden *kann*. Hinzu kommt, dass Sprachnutzer nicht perfekt sind und sich somit auch Daten in Korpora befinden, die die untersuchten sprachlichen Eigenschaften nicht angemessen wiedergeben. Solche Fälle sollten jedoch durch eine quantitative Auswertung der Ergebnisse herausgefiltert werden können.

Eine Korpusuntersuchung ist (nur) geeignet für Fragestellungen, die von klar **beobachtbaren Eigenschaften** eines linguistischen Phänomens abhängen (z.B. morphologische, syntaktische oder testbare semantische Eigenschaften, Kollokationen, Kookkurrenzen). Vor Beginn einer Korpusuntersuchung muss daher zunächst ermittelt werden, welche Aspekte der Fragestellung sich beobachten lassen.

> **Beispiel 1: *flink* und *flott***
> Ich möchte wissen, ob es einen Unterschied in den Bedeutungen der Adjektive *flink* und *flott* gibt. Was kann ich konkret beobachten? – In welchem Kontext die jeweiligen Adjektive auftreten: Wenn es einen Bedeutungsunterschied zwischen zwei Adjektiven gibt (es sich also nicht um totale Synonymie handelt), sind sie nicht (immer) austauschbar, treten also in unterschiedlichen Kontexten auf.
> In unterschiedlichen Kontexten treten zwei Adjektive beispielsweise dann auf, wenn sie unterschiedliche Nomina modifizieren. Diese Nomina können wiederum von einem unterschiedlichen semantischen Typ sein (z.B. Objekt, Ereignis, Zustand, Trope). So kann sich das Adjektiv *flott* z.B. auf *Tempo* beziehen, das Adjektiv *flink* hingegen nicht:
> (1) flottes Tempo
> (2) ?? flinkes Tempo

Nun sind Eigenschaften jedoch nicht von sich aus beobachtbar, sondern müssen erst dazu gemacht werden – und zwar nicht nur in der Linguistik: Das Gewicht eines Körpers beispielsweise kann ihm nicht einfach angesehen werden, sondern muss mithilfe einer Waage beobachtbar gemacht werden. Ebenso kann z.B. die grammatische Funktion SUBJEKT je nach Sprache erst anhand der Wortstellung (z.B. Englisch, Französisch) oder des Kasus (z.B. Nominativ im Deutschen) ermittelt werden. Es muss also nicht nur **festgelegt** werden, ***was* beobachtet werden kann**, sondern auch, ***wie*.**

Diese Festlegung nennt man auch **Operationalisierung**. Sie bildet die Grundlage für das weitere Vorgehen in der Korpusuntersuchung. Operationalisiert wird, wie das zu untersuchende linguistische **Phänomen** definiert wird, welche **Eigenschaften** beobachtet werden können und welche konkreten **Ausprägungen** diese Eigenschaften in den Korpusdaten aufweisen können.

Beispiel 2: *aber* und *nicht*
In der Literatur wird ein Zusammenhang zwischen Kontrast und Negation angenommen: Ein positiver Satz wird mit einem negativen verbunden („x, aber nicht y“). Ich möchte diesen Zusammenhang im Korpus untersuchen und lege dazu fest, dass das **Phänomen** KONTRAST mithilfe des kontrastiven Konnektors *aber* und das Phänomen NEGATION mithilfe der Negationspartikel *nicht* betrachtet werden soll.
Eigenschaften dieser Phänomene, die ich konkret beobachten kann, sind z.B. die relative Position von *nicht* zu *aber* und der Skopus von *nicht*.
Diese Eigenschaften können verschiedene **Ausprägungen** haben: Die relative Position kann „x, aber nicht y“ oder „nicht x, aber y“ sein, der Skopus der Negationspartikel kann ein Attribut, ein Komplement, ein Adjunkt oder das Verb sein.
Indem ich mir die Verteilung dieser unterschiedlichen Ausprägungen in den Korpusdaten ansehe, kann ich Hypothesen über den Zusammenhang der beiden Phänomene KONTRAST und NEGATION ableiten.

Dem *was* und *wie* folgt das ***wo***: Je nach Fragestellung und Operationalisierung ist die nächste Frage, **welches Korpus** die Beobachtungen ermöglicht. Hier kann berücksichtigt werden, ob nach konkreten Kategorien oder Ausdrücken gesucht werden soll, ob das Korpus also zu jedem Wort morphosyntaktische Informationen enthalten („getaggt sein“) soll oder nicht, oder ob bestimmte regionale oder zeitliche Aspekte für die Suche relevant sind.

Für die Erstellung einer geeigneten Suchanfrage ist neben der Operationalisierung des Phänomens auch das Einhalten der systemeigenen **Anfragesyntax** wichtig. Die Kombination aus den gesuchten Ausdrücken und einer Reihe von Operatoren, die z.B. die Wortform oder den Abstand der Ausdrücke zueinander betreffen, ermöglicht eine weitaus präzisere Suche, als sie über beispielsweise Google möglich wäre.

ⓘ Je **präziser** die Suchanfrage formuliert ist, desto **weniger unerwünschte Treffer** müssen im Nachhinein manuell aussortiert werden!

Nachdem die Ergebnisse ausgewählt und exportiert wurden, folgt der Schritt der **Annotation**. Deren Grundlage bildet die Operationalisierung der Eigenschaften und derer Ausprägungen: Für jeden Treffer wird annotiert, welche Ausprägung der Eigenschaften jeweils vorliegt. In einem letzten Schritt werden die Ergebnisse der Annotation (also die Häufigkeit bzw. Verteilung der einzelnen Ausprägungen) schließlich **ausgewertet** und mit der ursprünglichen Fragestellung abgeglichen.

Abb. 1 Übersicht: Ablauf einer Korpusuntersuchung

Abb. 1 stellt die Eckpunkte für diesen Ablauf einer Korpusuntersuchung dar und zeigt auf, in welchen Kapiteln Besonderheiten und wichtige Überlegungen zu diesen Arbeitsschritten vorgestellt werden. Die beiden Beispielfälle werden dabei an passenden Stellen zur Erläuterung wieder aufgegriffen.

2. Die Wahl des Korpus

Die Korpora des IDS, die mithilfe des Tools COSMAS II durchsucht werden können, bestehen zum größten Teil aus dem Deutschen Referenzkorpus (DeReKo), der weltweit größten Sammlung deutschsprachiger Korpora. Diese Korpora sind je nach Textsorte, Erscheinungsdatum, Annotation und Recherchemöglichkeit in **verschiedene Archive** eingeteilt. Die jeweiligen Archive haben unterschiedliche Eigenschaften, die für die Fragestellung relevant sein können. Somit ist die Wahl des passenden Archivs **von der Fragestellung abhängig**.

Die wichtigsten Archive sind *W* – Archiv der geschriebenen Sprache, das Wikipedia-Archiv *WP*, die Archive *TAGGED-C/2* und *TAGGED-T/2* sowie das Archiv *HIST* der historischen Korpora.

W – Archiv der geschriebenen Sprache

Das Archiv W ist das größte Archiv, umfasst die größte Bandbreite an Textsorten mit Texten vom 18. Jahrhundert bis heute. Es enthält u.a.:

- Zeitungskorpora aus Deutschland, Österreich und der Schweiz
- Korpora thematischer Zeitschriften (z.B. Computerzeitung, ZEIT Wissen)
- Chat-Korpora und Wikipedia-Artikel
- belletristische Korpora, Plenarprotokolle und politische Reden

Dieses Archiv eignet sich besonders für eine repräsentative Recherche, da es nicht nur sehr umfangreich, sondern auch ausgewogen und ausbalanciert ist. D.h. es enthält eine große Bandbreite an Textsorten, deren Anteil im Gesamtkorpus zudem ungefähr ihrem „realen“ Anteil entspricht. Außerdem wird es laufend erweitert, d.h. die Daten, die man bei einer Recherche erhält, sind aktuell.

Aufgrund der sehr großen Menge an Treffern, die alle einzeln annotiert werden müssen (siehe 4.2.), empfiehlt es sich, eine Zufallsauswahl der Belege in gewünschter Menge zu verwenden.

Wikipedia-Archiv *WP*

Wikipedia-Artikel sind auch bereits im Archiv *W* enthalten. In diesem Archiv sind aber zusätzlich auch die Wikipedia-Diskussionen aus den Jahren 2013, 2015 und 2017 enthalten. Das Archiv *WP* eignet sich besonders für die Untersuchung von in der Internetkommunikation verwendeten Ausdrücken.

Archive *TAGGED-C/2* und *TAGGED-T/2*

In diesen Archiven sind ca. 40% der Texte bis 2009 (*TAGGED-C bzw. TAGGED-T*) und ab 2010 (*TAGGED-C2 bzw. TAGGED-T2*) aus dem Archiv *W* enthalten, die hier zusätzlich morphosyntaktisch annotiert wurden. Zu jedem Wort sind also abgesehen vom Primärtext noch Metadaten morphosyntaktischer Natur, sogenannte *Tags*, gespeichert. So kann in diesen Archiven mithilfe des MORPH-Assistenten nicht nur nach konkreten Ausdrücken gesucht werden, sondern allgemein nach bestimmten Wortarten und Wortformen (dies ist im Archiv *W* nicht möglich).

> **Beispiel 1: *flink* und *flott***
> Für die Untersuchung eines Bedeutungsunterschiedes zwischen *flink* und *flott* ist entscheidend, mit welchen Nomina die Adjektive auftreten – hier empfiehlt sich also die Wahl eines getaggten Korpusarchivs, in dem gezielt nach der Kombination aus attributivem Adjektiv und Nomen gesucht werden kann, z.B. *TAGGED-T2*.

Ein weiterer Vorteil ist, dass die Tags ebenfalls exportiert werden können. So kann die Auswahl dieser Archive auch sinnvoll sein, wenn beispielsweise die Distribution verschiedener Wortarten um einen Ausdruck herum untersucht werden soll. Aus diesen Gründen sind **getaggte Korpora** allgemein **zu bevorzugen**.

Der Unterschied zwischen den beiden Archiv-Arten *T* bzw. *C* besteht im jeweils für die automatische Annotation verwendeten Tagger-Set, *stts* bzw. *Connexor*. Hier stehen unterschiedliche Wortarten, -unterarten und Flexionsklassen zur Verfügung, wodurch die Wahl des Archivs stark von der Fragestellung abhängig ist (mehr dazu siehe 3.2.). Zu beachten ist, dass die Annotation aufgrund des großen Umfangs jedoch rein maschinell erfolgt, wodurch die Tags teilweise fehlerbehaftet sein können, was eine manuelle Überprüfung der Treffer unabdingbar macht (siehe 4.1.).

Archiv *HIST* der historischen Korpora

HIST enthält Texte von ca. 1650 bis 1962, so u.a. Texte von Marx & Engels, Goethe, Brüder Grimm und historische Zeitungstexte. Eine Suche in diesem Archiv ist sinnvoll für diachrone Studien oder die Untersuchung eines sprachlichen Phänomens in einer früheren Sprachstufe. Eine Recherche im Archiv *HIST* kann auch vergleichend einer Suche im Archiv *W* (mit aktuellen Texten) gegenübergestellt werden.

Nach der Auswahl eines Archivs wird ein **Korpus** ausgewählt. In allen Archiven ist bereits eine Korpusauswahl vordefiniert und voreingestellt. Im Archiv *W* ist das zum Beispiel das Korpus *W-öffentlich*, mit dem alle öffentlich zugänglichen Texte des Archivs *W* durchsucht werden. In Abhängigkeit zur Fragestellung kann es jedoch auch sinnvoll sein, nur einzelne relevante Korpora auszuwählen oder einzelne Korpora von der Suche auszuschließen.

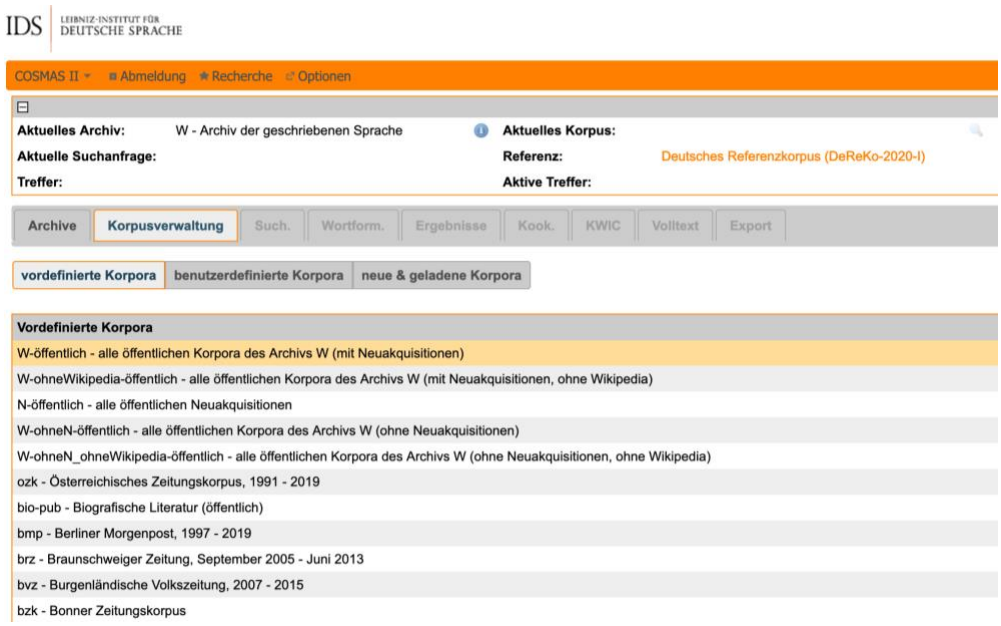


Abb. 2 Archiv- und Korpusauswahl in COSMAS II

Die Auswahl einzelner Korpora im Archiv *W* ermöglicht einem zum Beispiel regionale/dialektale Unterschiede zu erforschen, etwa indem man deutsche und österreichische Zeitungskorpora miteinander vergleicht. Umgangssprachliche bzw. in der Internetkommunikation verwendete Ausdrücke können beispielsweise im Dortmunder Chatkorpus 2.1 (Korpus *dzk*) untersucht werden. Auch im Archiv *HIST* können einzelne Korpora ausgewählt werden, wie etwa Goethes Werke im Korpus *goe*.

Beispiel 2: *aber* und *nicht*
Für die Untersuchung des Zusammenhangs von Kontrast und Negation ist keine konkrete Einschränkung nötig. Um sehr umgangssprachliche Verwendungsweisen auszuschließen, könnte man jedoch Wikipedia-Artikel von der Suche ausgrenzen. Somit wäre die Wahl des Archivs *W* und daraufhin des Korpus *W-ohneWikipedia-öffentlich* eine passende Option.

3. Die Suchanfrage in COSMAS II

Abb. 3 zeigt das Suchfeld über COSMAS II$_{web}$: Im Header ist zunächst zu sehen, welches Archiv und welche Korpora ausgewählt wurden. Im Eingabefeld wird die Suchanfrage entsprechend der Anfragesyntax formuliert. Die Suchanfrage sollte **so präzise wie möglich** formuliert sein, damit die manuelle Überprüfung und Aussortierung ungeeigneter Treffer hinterher minimiert werden kann.

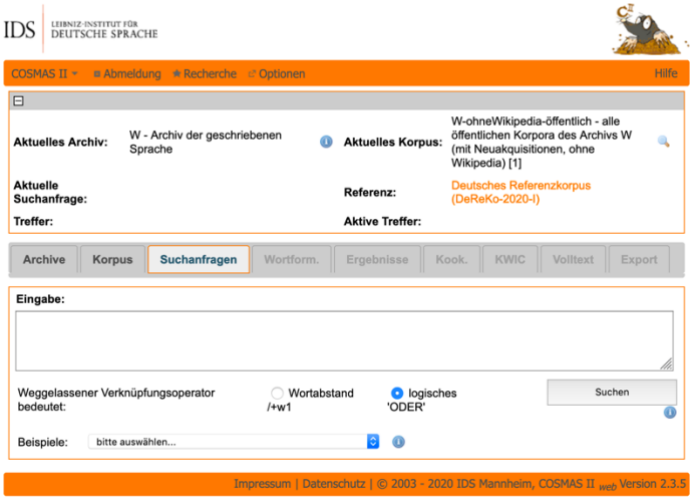


Abb. 3 Suchfeld über cosmas II

Die Suchanfrage kann bereits aus einem **konkreten Ausdruck** bestehen, z.B. *Universität*. Sind mehrere Ausdrücke zu suchen, z.B. *Universität Tübingen*, wird die Angabe unterhalb des Suchfelds („Weggelassener Verknüpfungsoperator bedeutet:“) relevant: Ist die Option links, „Wortabstand /+w1“ ausgewählt, liefert COSMAS Belege für beide Wörter unmittelbar hintereinander in der angegebenen Reihenfolge. Satzzeichen können allerdings dazwischenstehen. Diese Variante kann z.B. für die Suche nach konkreten Phrasen genutzt werden. Ist die Option rechts, „logisches ‘ODER‘“, ausgewählt, liefert COSMAS Belege, in denen entweder das Wort *Universität* oder das Wort *Tübingen* oder beide in beliebiger Reihenfolge innerhalb eines Textes auftauchen. Diese Variante kann z.B. für die Suche ähnlicher Begriffe verwendet werden.

Bei einer solchen „einfachen“ Suche wird jeweils die konkret angegebene Wortform gesucht, also z.B. nur *Universität*, nicht aber *Universitäten*. Die Suchanfrage über COSMAS II kann aber noch viel mehr. Dazu wird eine **Anfragesyntax** verwendet, deren gebräuchlichste Operatoren in 3.1. in ihren Grundzügen erläutert werden. Die Anfragesyntax ist in allen über COSMAS zugänglichen Korpora anwendbar, auch in den getaggten. Hier kommen dann noch Suchoptionen nach morphosyntaktischen Tags ergänzend hinzu, siehe 3.2.

> ⓘ Wenn eine Suche zunächst keine Ergebnisse liefert, heißt das noch lange nicht, dass es das Gesuchte im Korpus nicht gibt. Es gilt, nochmals und ggf. mehrmals einen Blick auf die Formulierung der Anfrage zu werfen und diese anzupassen. Diese **Optimierung der Suchanfrage** ist ein Prozess, der notiert werden sollte: Welche Formulierung der Suchanfrage ergab welche Art von Treffern? Was war das Problem damit? Mit welcher neuen Formulierung soll dieses Problem vermieden werden?

3.1. Allgemeine Anfragesyntax in COSMAS II

Für die komplexe Suche mit mehreren Suchbegriffen bzw. Syntagmen stehen neben den Standardverknüpfungsoperatoren Wortabstand /+w1 und logisches 'ODER' eine Reihe weiterer **Operatoren** zur Verfügung, die zusammen mit den Suchbegriffen in die Suchmaske eingegeben werden.

Die wichtigsten Operatoren werden im Folgenden erläutert. Dazu gehören Operatoren, die die Wortform des gesuchten Ausdrucks betreffen, und sogenannte Abstandsoperatoren, die bei mehreren Suchbegriffen oder Eigenschaften relevant sind. Auch Operatoren, die die Position des Suchbegriffs im Satz einschränken, können hilfreich sein. Schließlich gibt es ein paar Dinge für „logische“ Verknüpfungsoperatoren und Satzzeichen zu beachten. Die vollständige Liste sowie weitere Anwendungsbeispiele können online abgerufen werden.

Wortformoperatoren

Mit dem **Grundformoperator &** sucht man nach **Flexions- und Wortbildungsformen** einer Grund- bzw. Nennform. Die Standardeinstellung ist dabei die Suche nach Flexionsformen. D.h. immer, wenn *&* unmittelbar vor einem Ausdruck steht, werden alle vorhandenen Flexionsformen dieses Ausdrucks gesucht. So wird bei der Eingabe `&Universität` neben der Nennform *Universität* auch die Pluralform *Universitäten* gesucht.

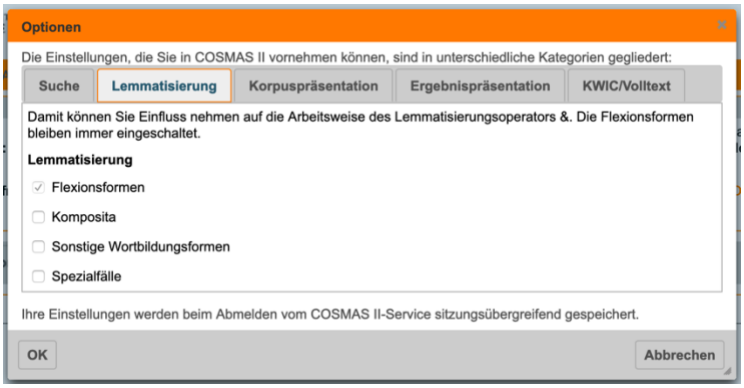


Abb. 4 Optionen > Lemmatisierung

Der Grundformoperator kann auch für die Suche nach **Komposita und Derivaten** verwendet werden. Hierzu muss unter *Optionen > Lemmatisierung* die Standardeinstellung von *Flexionsformen* auf die gewünschten Wortbildungsformen geändert werden (siehe Abb. 4). Wird dort *sonstige Wortbildungsformen* ausgewählt, kann so beispielsweise mit `&-heit` nach allen mit dem Suffix *-heit* gebildeten Wörtern gesucht werden (in diesem Fall 7.260 Wortformen von *Aalglattheit* bis *Zwiegespaltenheit*).

> ⓘ Bei trennbaren Verben (Partikelverben) findet COSMAS mit dem Grundformoperator **nur nicht-getrennte Verbformen** (Eingabe: `&ankommen` → *ankommt*, *ankam*, *ankommend*, …). Möchte man nach Formen suchen, bei denen die Partikel vom Verbstamm getrennt ist (z.B. *kommt an*), muss dies separat mithilfe eines Abstandsoperators (siehe unten) getan werden (Eingabe: `&kommen /s0 an`). Aber Vorsicht: Ohne weitere morphosyntaktische Informationen, die nur über getaggte Korpora zugänglich sind, unterscheidet COSMAS an dieser Stelle nicht, ob es sich bei *an* um eine Verbpartikel oder eine Präposition handelt!

Weitere Operatoren, die die Wortform des Suchbegriffs betreffen, sind die **Platzhalteroperatoren**. Diese sind nicht in ihrer Position im gesuchten Ausdruck beschränkt und umfassen je nach Symbol eine unterschiedliche Menge an Zeichen. Hier eine **Übersicht der Wortformoperatoren** mit Beispielen:

Was?	Wozu?	Beispiel
`&`	Flexions-/Wortbildungsformen	`&Tisch` → *Tisch, Tische, Tisches*
`*`	Platzhalter für 0 bis n Zeichen	`Tisch*` → *Tisch, Tische, Tischtennis, Tischtuch, …*
`?`	Platzhalter für genau 1 Zeichen	`??ständig` → *beständig, anständig, zuständig, …*
`+`	Platzhalter für 0 oder 1 Zeichen	`++ständig` → *ständig, beständig, anständig, …*

> **Beispiel 1: *flink* und *flott***
> Untersucht werden die Nomina, die mit den attributiv verwendeten Adjektiven auftreten. Diese können natürlich nicht nur Unterschiede im zu annotierenden Typ des Nomens haben, sondern auch in Genus, Kasus und Numerus. Die Suche sollte daher die verschiedenen Flexionsformen der Adjektive einschließen, also `&flink` bzw. `&flott`.

Abstandsoperatoren

Abstandsoperatoren ermöglichen die Suche nach mehreren Begriffen, die in einem bestimmten maximalen Abstand zueinander vorkommen („**treffereinschließend**“) oder nicht vorkommen („**trefferausschließend**“). Die Abstände werden mit Zahlen angegeben und können auf die Wort-,

Satz- und Absatzebene angewendet werden. Somit bestehen die Abstandsoperatoren aus einem **Sonderzeichen** für trefferein- (/) bzw. -ausschließend (%), den **Buchstaben** w, s oder p für die entsprechende Ebene (also Wort, Satz oder Absatz) sowie einer **Zahl** für den **maximalen** Abstand. Ein minimaler Abstand kann mittels Doppelpunktes vor dem maximalen Abstand angezeigt werden.

Der Operator /w1 zwischen zwei Ausdrücken beispielsweise bedeutet, dass die beiden gesuchten Ausdrücke unmittelbar hintereinander auftreten sollen, /w2:4 dass die Wörter mindestens zwei, maximal aber vier Wörter auseinander auftreten. Der Operator %s0 wiederum bedeutet, dass die Ausdrücke nicht innerhalb eines Satzes auftreten. Wenn dies etwas genauer werden soll (z.B., dass die Ausdrücke in Folgesätzen auftreten sollen), empfiehlt sich die treffereinschließende Version /s1.

Zusätzlich kann noch die **Reihenfolge** der Ausdrücke im angegebenen Abstand zueinander angegeben werden, mit + für die in der Suchanfrage gegebene Reihenfolge und – für die entgegengesetzte. Hier eine **Übersicht mit Beispielen**:

Wo?	**Wie?**	**Beispiele**
Wortebene	/w…	`rot %+w1 gestreift` → *rot ~~gestreift~~*
	%w…	`rot /+w3 gelb` → *rot gelb, rot und gelb, rot und nicht gelb, …*
		`rot /+w2:3 gelb` → *rot und gelb, rot und nicht gelb, …*
Satzebene	/s…	`rot /s0 gelb` → *rot* und *gelb* im selben Satz
	%s…	`rot %s0 gelb` → *rot* und *gelb* nicht im selben Satz
		`rot /s1 gelb` → *rot* und *gelb* im selben Satz und in aufeinanderfolgenden Sätzen
Absatzebene	/p…	`erstens /+p1 zweitens` → *erstens* gefolgt von *zweitens* im selben oder darauffolgenden Absatz
	%p…	

Positionen im Satz

Auch für die nicht-getaggten Korpora in COSMAS II stehen Metadaten zur Verfügung, die in die Suche einbezogen werden können. Diese betreffen hauptsächlich die verschiedenen Ebenen der Textstruktur (Wort, Satz, Absatz, Überschrift; siehe Abstandsoperatoren). Dies kann dazu genutzt werden, die Position eines Ausdrucks, z.B. innerhalb eines Satzes, einzuschränken.

> ⓘ Ein **„Satz“** ist in COSMAS nicht als die linguistische Einheit zu verstehen, sondern maßgeblich das, was durch ein **satzbeendendes Zeichen** markiert ist (Punkt, Fragezeichen). Durch Kommata getrennte Hauptsätze oder die Kombination aus Haupt- und **Nebensätzen** werden also hier **nicht berücksichtigt**!

Um für Suchbegriffe eine bestimmte Position innerhalb des „Satzes“ festzulegen, wird der **Abstandsoperator /w0** mit einer **Satzpositionsangabe** kombiniert. So wie der Abstandsoperator /s0 („Abstand von Null Sätzen“) bedeutet, dass die Suchbegriffe innerhalb eines Satzes gesucht werden, heißt /w0 („Abstand von Null Wörtern“) so viel wie „im selben Wort“. Der Wortabstandsoperator /w0 kann jedoch nicht für die Kombination verschiedener zu suchender Ausdrücke verwendet werden (da es sich sonst um Komposita handeln würde – dazu siehe Wortformoperatoren oben). Stattdessen verbindet der Operator einen Ausdruck mit einer **Eigenschaft**.

Eigenschaften sind zwar hauptsächlich in den getaggten Korpora relevant, jedoch gibt es überall zwei Satzeigenschaften: <sa> bedeutet Satzanfang, <se> bedeutet Satzende. Die Suchanfrage `ohne /w0 <sa>` würde also alle Sätze ergeben, die mit *ohne* beginnen. Alternativ kann die Suchanfrage auch `ohne:sa` lauten. Natürlich steht auch hier die **trefferausschließende Option** zur Verfügung. Die Variante mit dem Wortabstandsoperator %w0 ergibt eine Suche des Ausdrucks überall, außer am Satzanfang bzw. -ende.

Eine weitere Variante ist die Verwendung des #**IN-Operators**, auf deren Beschreibung online verwiesen wird.

Logische Operatoren und Satzzeichen

Wie jede Computersprache verwendet auch die Anfragesyntax von COSMAS II die **logischen Verknüpfungsoperatoren** `und`, `oder` und `nicht`. Die Eingabe `Universität nicht Tübingen` sucht z.B. nach dem Ausdruck *Universität* ohne *Tübingen* (also z.B. *Universität Stuttgart, Universität, Universität zu Köln*). Mithilfe von Klammerungen ist auch die Kombination mehrerer Verknüpfungsoperatoren möglich (vgl. Aussagenlogik): `(Universität und Tübingen) oder (Universität und Stuttgart).`

> ⓘ Die Suche nach einem Wort, das wie *und, oder* und *nicht* in der COSMAS-Anfragesyntax ein **Operator** ist, muss mithilfe von **Anführungszeichen** geschehen. Soll also beispielsweise nach der Konjunktion *und* gesucht werden, muss die Suchanfrage `"und"` lauten.

Da **Anführungszeichen** in COSMAS II zwischen Operatoren und zu suchenden Ausdrücken unterscheiden, haben diese auch eine Art Operatorfunktion. Wenn konkret nach Anführungszeichen gesucht werden soll, die Anführungszeichen also von COSMAS als Suchbegriff verstanden werden sollen, müssen sie mit einem Backslash (\) vorangestellt werden. Die Suche `\" /+w1 (&sagen oder &antworten)` eignet sich beispielsweise für die Suche nach direkter Rede gefolgt von (flektierten Formen der) redeeinleitenden Verben *sagen* oder *antworten*. Das Komma, das üblicherweise zwischen direkter Rede und einleitendem Verb steht, wird bei der Suche ignoriert.

Um **Satzzeichen** gezielt zu suchen oder von der Suche auszuschließen, wird wieder der **Wortabstandsoperator** eingesetzt. Soll beispielsweise ein Komma zwischen zwei Ausdrücken ausgeschlossen werden, muss die Eingabe `%w0 ,` lauten (z.B. `rot %w0 , grün`). Die Suche nach einem Komma geschieht mit der treffereinschließenden Variante `/w0 ,` für die Suche nach einem Komma nach dem gesuchten Ausdruck bzw. `, /+w1` für die Suche nach einem Komma vor dem Ausdruck. Andere Satzzeichen müssen, sofern sie auch eine Operatorfunktion in COSMAS haben, zusätzlich mit Anführungszeichen gesucht werden, z.B. `"?"`.

Da ein **Bindestrich** in COSMAS II keine Operatorfunktion einnimmt, kann nach diesem ganz ohne Anführungszeichen gesucht werden. In Verbindung mit den bisher erläuterten Operatoren können interessante Suchanfragen formuliert werden, wie beispielsweise `- /+w1:1,s0 "und"`, das eine Suche nach Koordinationen von Komposita mit gleichem Kopfwort ermöglicht, wie *Wahl- und Stimmrecht* oder *Plan- und Baugesuchsunterlagen*.

> **Beispiel 2: *aber* und *nicht***
> Wie müssen wir die Suchanfrage formulieren, um den Zusammenhang von Kontrast und Negation zu untersuchen? Wir wissen nun:
> – dass COSMAS II nur satzbeendende Zeichen als Satzgrenze ansieht. Um mit Kommata verbundene Hauptsätze („x, aber nicht y“) untersuchen zu können, sollte also ein Komma in die Suche einbezogen werden. → `, /+w1 aber`
> *Aber* und *nicht* sollten dann in freier Reihenfolge im selben „Satz“ auftreten. → `/s0`
> – dass die Suche nach der Negationspartikel *nicht* mithilfe von Anführungszeichen geschehen muss, da das Wort in der Anfragesprache von COSMAS II eine Operatorfunktion einnimmt. → `"nicht"`
> Die vollständige Suchanfrage (inklusive Komma) lautet also:
> **`(, /+w1 aber) /s0 "nicht"`**

3.2. Morphosyntaktische Tags

Wie in Abschnitt 2 erläutert, stehen in COSMAS II mit den Archiven *TAGGED-C/2 bzw. TAGGED-T/2* Korpora zur Verfügung, die maschinell **mit morphosyntaktischen Metadaten annotiert** wurden. Somit kann hier die bisher erläuterte Anfragesyntax um die Suche nach konkreten morphologischen Eigenschaften wie **Wortart** oder **Flexionsklasse** ergänzt werden.

Diese erfragbaren *Tags* können entweder durch ihr jeweiliges Kürzel in das Suchfeld eingegeben (z.B. `MORPH(N)` für Nomina) oder mithilfe des **MORPH-Assistenten** (siehe Abb. 5) per Dropdown-Menü ausgewählt werden. Dieser übersetzt die ausgewählten Kategorien dann in die Anfragesyntax.

Welche konkreten Wortarten, Unterarten und Flexionsklassen jeweils zur Verfügung stehen, **variiert zwischen den beiden Archiv-Arten *T* bzw. *C***.

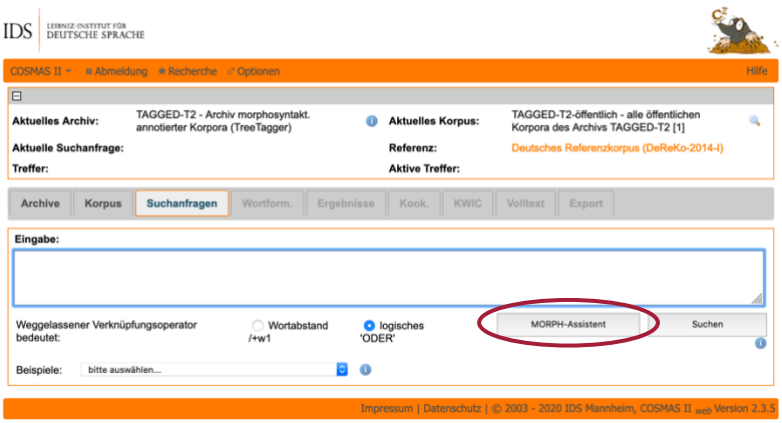


Abb. 5 MORPH-Assistenz unter Suchfeld

Korpora in den Archiven ***TAGGED-T* bzw. *TAGGED-T2*** verwenden das Tagger-Set namens **stts** (*Stuttgart-Tübingen-Tagset*). Der sich hier öffnende MORPH-Assistent erlaubt beispielsweise im Reiter *Verben* die Auswahl der Flexionsklassen *finit* (mit/ohne Imperativ), *Infinitiv, Partizip Perfekt* oder *beliebig* (Abb. 6 links). Anschließend kann die Verbklasse festgelegt werden (Abb. 6 rechts), wobei einzelne Klassen ausgewählt (+ links) bzw. ausgeschlossen (− links) werden können.

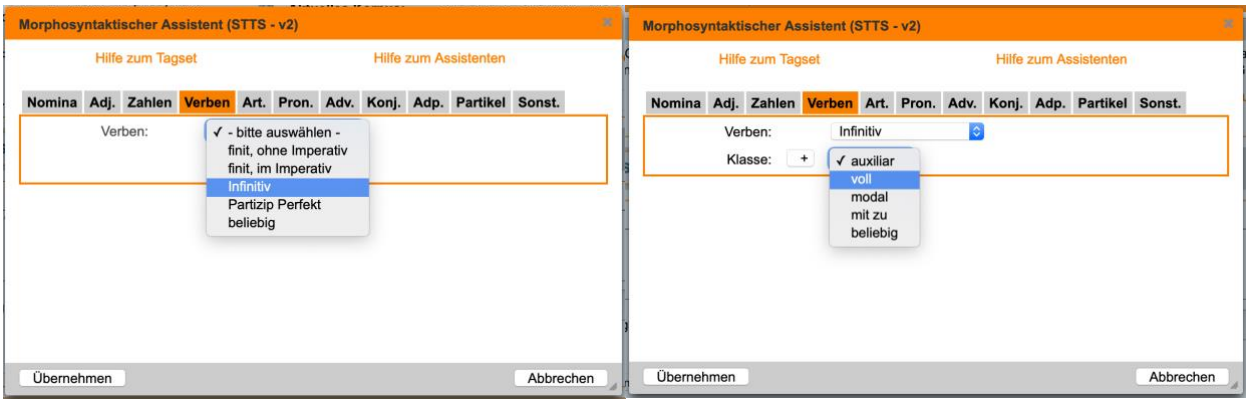


Abb. 6 MORPH-Assistent für stts (Verben)

Die Auswahl der Kategorien *Infinitiv* und *Vollverb* (eingeschlossen) übersetzt sich bei Übernahme im Suchfeld in den Suchbefehl `MORPH(VRB inf v)`. Ohne weitere Ergänzungen könnten mit dieser Suche also alle in den Korpora vorkommenden Vollverben im Infinitiv gesucht werden.

Korpora in den Archiven ***TAGGED-C* bzw. *TAGGED-C2*** verwenden das Tagset **CONNEXOR**. Für Verben können in dem sich hier öffnenden MORPH-Assistenten beispielsweise noch Angaben zu Modus (*Indikativ, Imperativ, Konjunktiv*), Verbform (*finite Form, Infinitiv, Partizip I* und *II*) und Tempus (*Präsens* und *Präteritum/Perfekt*) ausgewählt werden. Auch hier können die Subkategorien in die Suche einbezogen (+) oder von ihr ausgeschlossen werden (−) oder die allgemeine Option „beliebig“ ausgewählt werden (siehe Abb. 7).

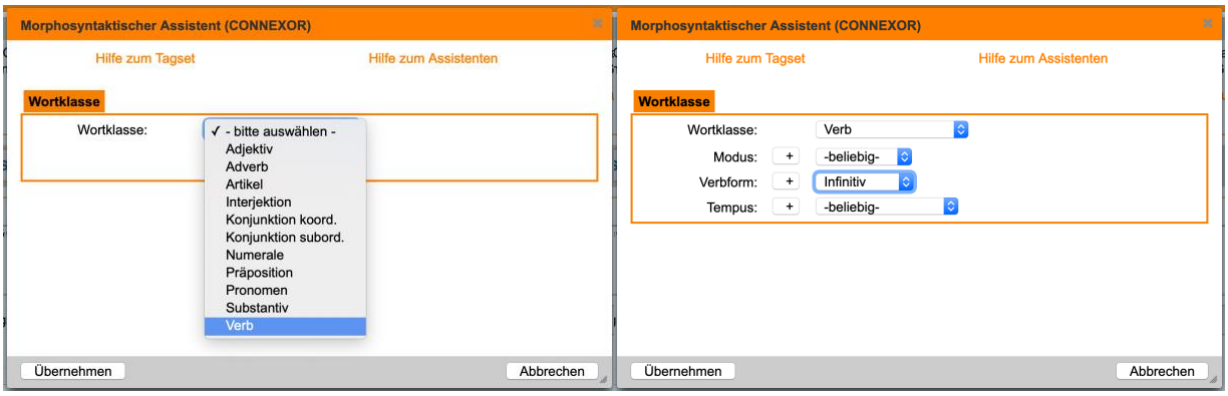


Abb. 7 MORPH-Assistenz für CONNEXOR (Verben)

Die Suche nach Verben im Infinitiv in diesem Tag-Set wird als `MORPH(V INF)` übersetzt, wobei allerdings im Unterschied zu der Suchanfrage in den *T*-Korpora Hilfs- und Modalverben nicht von der Suche ausgeschlossen werden.

> ⓘ Die Wahl der jeweiligen Archive *T* oder *C* ist abhängig von der Fragestellung bzw. den Bedürfnissen, welche Kategorie wie detailliert gesucht werden soll. Es ist nicht der Fall, dass dabei generell ein Tag-Set detaillierter ist als das andere, sie sind lediglich in unterschiedlichen Projekten entstanden und verfolgen somit **unterschiedliche Zwecke**. Es empfiehlt sich, **vor Auswahl des Archivs** die jeweiligen verfügbaren Kategorien zu betrachten. Diese sind online jeweils für stts und CONNEXOR in einer **Übersicht** aufgelistet. (Wenn sich für die Fragestellung beide Tagsets gleichermaßen eignen, nehmen Sie doch das „hauseigene“ stts. ☺)

Die per MORPH-Assistent erstellten Suchbefehle können **mit** den in 3.1. erläuterten **Operatoren kombiniert** werden. Wie dort gezeigt, ist vor allem der **Wortabstandsoperator `/w0`** relevant, der einen Ausdruck mit einer Eigenschaft verbindet. Soll ein gesuchter Ausdruck Träger der Eigenschaft sein, muss die treffereinschließende Variante `/w0` verwendet werden, bei auszuschließenden Eigenschaften dementsprechend die Variante `%w0`. Soll beispielsweise bei der Suche nach dem Verb *laufen* in den *C*-Korpora die Imperativ-Form ausgeschlossen werden, ergibt sich gemeinsam mit dem MORPH-Assistenten die Suchanfrage `&laufen %w0 MORPH(V -IMP)` (gesucht werden alle Flexionsformen des Verbs *laufen*, außer jene im Imperativ).

Auch die Wortformoperatoren können selbstverständlich mit dem MORPH-Assistenten kombiniert werden. So kann beispielsweise mithilfe der Anfrage `MORPH(VRB inf v) /w0 &ent-` in den *T*-Korpora nach allen mit dem Präfix *ent-* gebildeten Verben im Infinitiv gesucht werden.

> ⓘ Durch die rein maschinelle Annotation der morphosyntaktischen Eigenschaften sind die **Tags fehlerbehaftet**. So kann es beispielsweise vorkommen, dass der Tagger ein Verb wie *entwenden* als Infinitiv markiert hat, obwohl es sich im konkreten Treffer um die gleichlautende Form der 3.Pers.Pl. handelt. Solche Fälle müssen vor dem Export **manuell überprüft und ausgeschlossen** werden (siehe 4.1.).

Beispiel 1: *flink* und *flott*
Wie muss die Suchanfrage nach der Kombination der Adjektive *flink* bzw. *flott* mit darauffolgenden Nomina aussehen? Wir wissen nun:
– dass die verschiedenen Flexionsformen der Adjektive mithilfe des Grundformoperators gesucht werden können. → `&flink` bzw. `&flott`
– dass ein Abstandsoperator ein darauffolgendes Wort sucht. → `/+w1`
– dass mithilfe des MORPH-Assistenten in getaggten Korpora nach einer konkreten Wortart wie Nomen gesucht werden kann. → `MORPH(N)` (in *TAGGED-T/2*)
Die vollständigen Suchanfragen lauten also: **`&flink`** bzw. **`&flott /+w1 MORPH(N)`**

4. Die Ergebnisse

Vor Ausführung der fertigen Suchanfrage können unter ***Optionen > Suche*** nützliche **Voreinstellungen** getroffen werden. Je nach Häufigkeit der gesuchten Ausdrücke und Umfang der ausgewählten Korpora kann es beispielsweise ratsam sein, die Ergebnismenge bereits vor der Durchführung der Suche auf eine **Zufallsauswahl** mit festgelegter Ergebniszahl zu begrenzen.

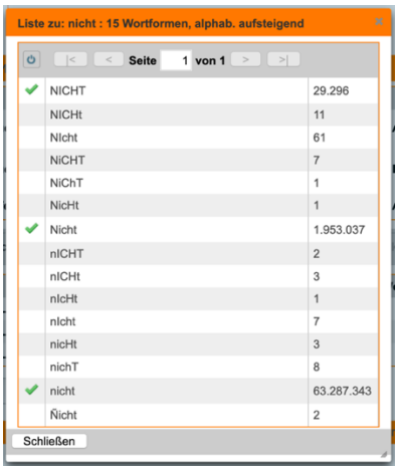


Abb. 8 Wortformen-Liste vor Suche

Bevor COSMAS II die teilweise ein paar Minuten dauernde Suche tatsächlich ausführt, wird zunächst an der **Oberfläche** nach verschiedenen **Wortformen** gesucht.

Für die Suche nach der Negationspartikel *nicht* (`"nicht"`) werden im Korpus *W-öffentlich* beispielsweise 15 Wortformen gefunden. Wie Abb. 8 zeigt, sind manche davon merkwürdig, z.B. *Ñicht* ganz unten. Auch wenn diese nicht sehr häufig sind (vgl. rechte Spalte), könnten sie bei einer Suche mit Zufallsauswahl durchaus in den Ergebnissen auftauchen und müssten dann manuell aussortiert werden. Um das zu vermeiden, können in der linken Spalte per Mausklick nur jene **Wortformen ausgewählt** werden, die einbezogen werden sollen.

> ⓘ Die **Überprüfung** der verschiedenen Wortformen kann bereits einen Hinweis liefern, ob die Suchanfrage zumindest dahingehend geglückt ist!

Erst nach der möglichen Kontrolle der verfügbaren Wortformen wird die Suche mittels Klicks auf den ***Ergebnisse*-Button** auch tatsächlich ausgeführt. Je nach Voreinstellungen werden die Ergebnisse zunächst in der Korpusansicht präsentiert, in der angezeigt wird, wie viele Treffer in welchem Korpus gefunden wurden. Hier können sie z.B. nach Jahr sortiert werden und pro Korpus einzeln angesehen werden.

Die Gesamtheit der Treffer kann unter dem Reiter **KWIC** bzw. **Volltext** eingesehen werden. In der KWIC-Ansicht werden die gesuchten Ausdrücke in ihrem unmittelbaren Kontext farbig hervorgehoben („key word in context“, siehe auch Abb. 9), die Volltext-Ansicht zeigt einen größeren Kontext. Wie viel Kontext pro Ansicht gezeigt wird, kann ebenfalls unter *Optionen* eingestellt werden.

Auch an dieser Stelle ist noch einmal eine Überprüfung möglich, ob die formulierte Suchanfrage die gewünschte Art von Treffern liefert.

4.1. Export der Ergebnisse

Um die Daten im Anschluss an die Suche weiter zu verwenden, müssen sie exportiert werden. Dazu können entweder alle Treffer bzw. eine Zufallsauswahl verwendet oder unerwünschte Treffer bereits in der KWIC- bzw. Volltext-Ansicht **manuell aussortiert** werden.

In Abb. 9 ist die KWIC-Ansicht für die Suche unseres Beispiels 1. Hier könnten z.B. Treffer 10 und 11 aussortiert werden. Die Treffer 2, 4 und 5 zeigen die gleiche Kombination *flinke Finger*. Da es um die Untersuchung potentiell unterschiedlicher Typen von Nomen geht,

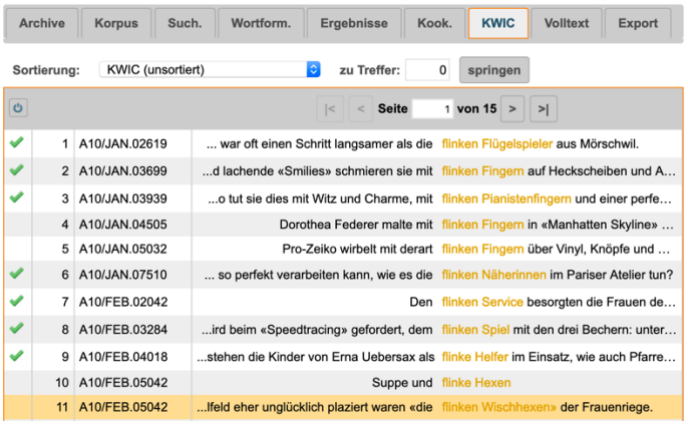


Abb. 9 Treffer in KWIC-Ansicht flink *+ Nomen*

könnten solche gleichlautenden Phrasen aussortiert werden. Aber Vorsicht: Wenn sich die **Auswertung** auf die Häufigkeit der verschiedenen Nomen-Typen beziehen soll, würde eine Aussortierung dieser Beispiele diese **verfälschen**!

> ⓘ Der **Ausschluss von Treffern** darf **nicht willkürlich** sein! Jeder vorgenommene Ausschluss muss in Hinblick auf die Fragestellung begründbar sein und in der wissenschaftlichen Arbeit erläutert. Mögliche Gründe sind beispielsweise ambige oder unverständliche Sätze oder Treffer aus einer bestimmten irrelevanten Kategorie (z.B. stark metaphorische Verwendungen). Wenn letztere zu häufig vorkommen, kann sich unter Umständen die Überarbeitung der Suchanfrage lohnen. Die Ausschluss-Kriterien müssen **konsequent** und einheitlich angewendet werden.

Beispiel 1: *flink* und *flott*
Die KWIC-Ansicht der Ergebnisse (Suchanfragen `&flink` bzw. `&flott /+w1 MORPH(N)` im Korpus *TAGGED-T2-öffentlich*) zeigt, dass diese Suchanfragen präzise genug formuliert sind. Manuell aussortiert werden können folgende Arten von Treffern:
(3) *Flinke „Bluegrass“- oder Jazzkompositionen* (Beschränkung auf nur ein Nomen)
(4) *die „Flinken Spitzen“ der Kyffhäuser Kameradschaft Werlaburgdorf* (Eigenname)

Unter dem Reiter *Export* können nun die Einstellungen vorgenommen werden, welche Treffer wie exportiert werden sollen (siehe Abb. 10). Dies betrifft zunächst allgemeine Informationen, wie Dateiname und **Format**. ASCII heißt, es wird eine .txt-Datei gespeichert. RTF („Rich-Text-Format“) bietet etwas mehr Raum für das Format und kann mit verschiedenen Programmen geöffnet werden).

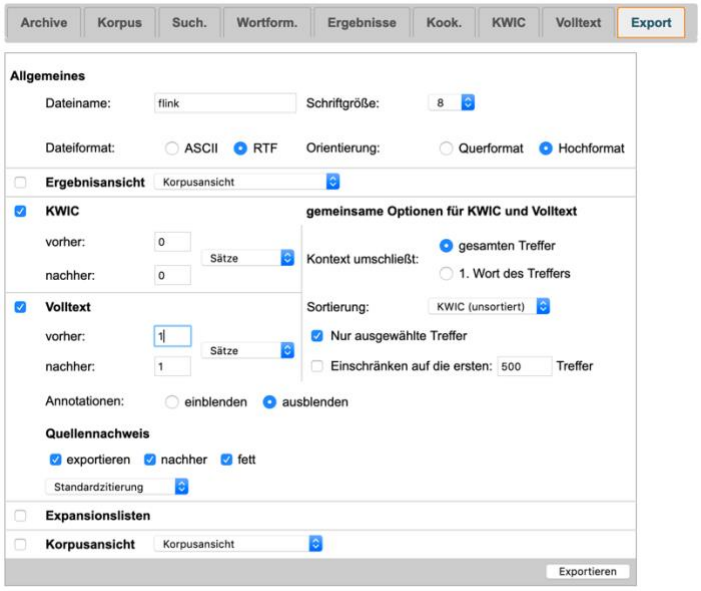


Abb. 10 Export-Einstellungen

Weiterhin kann eingestellt werden, wie viel **Kontext** rund um die Suchbegriffe exportiert werden soll. Auch das variiert nach Fragestellung. Da sich eventuell bei der Annotation herausstellen kann, dass für die Interpretation der Daten größere Kontexte nötig sind als zunächst gedacht, empfiehlt sich der Export eines größeren Kontexts.

In Korpora aus den getaggten Archiven steht die Möglichkeit zur Verfügung, die **Tags** ebenfalls zu exportieren. Das kann in Untersuchungen sinnvoll sein, bei denen die Distribution verschiedener Wortarten um einen Ausdruck herum relevant sind.

> ⓘ

Wichtig ist, dass in jedem Fall der **Quellennachweis** exportiert wird. Jeder Treffer erhält ein Kürzel (vgl. (5) und (6) im Beispielkasten), mithilfe dessen er zurückverfolgt werden kann. Das ist für die Reliabilität der Korpusuntersuchung nötig. Schließlich kann eingestellt werden, ob die Ergebnisse in irgendeiner Form sortiert sein sollen und ob die manuell ausgewählten Treffer (vgl. Abb. 9) oder eine zufällige Auswahl aus den Ergebnissen verwendet werden soll.

Beispiel 2: *aber* und *nicht*
Da der Zusammenhang von *aber* und *nicht* in ganzen Sätzen untersucht werden muss, sollten die Ergebnisse der Suchanfrage `(, /+w1 aber) /s0 "nicht"` in der Volltext-Ansicht überprüft werden. Eine zusätzliche manuelle Aussortierung der Treffer vor

Export ist hier aus (mindestens) zwei Gründen sinnvoll und nötig: Die in COSMAS verwendete Definition von „Satz“ ergibt auszusortierende Treffer wie in (5), bei dem die Negation erst im Folgesatz auftaucht, und der Skopus der Negation kann höher sein als das Kontrastziel von *aber* wie in (6).

(5) In Westeuropa seien es auch heute noch vielfach Frauen, **aber** auch wirtschaftlich Schwächere und Fremde, die an den Rand der Gesellschaft gedrängt werden – und mit «Fremde» meinte Barbara Scheffer durchaus **nicht** nur Ausländerinnen. (A97/APR.00754 St. Galler Tagblatt, 26.04.1997)

(6) dass es **nicht** möglich sei, den Kredit zwar anzunehmen, **aber** das Projekt abzulehnen (A97/APR.00046 St. Galler Tagblatt, 23.04.1997)

4.2. Annotation der Ergebnisse

Nach dem Finden der Belege kommt die eigentliche Interpretationsarbeit, die Annotation der Daten. Hierbei werden die zu Beginn der Korpusuntersuchung operationalisierten Eigenschaften bzw. deren Ausprägungen auf die konkreten Belege angewendet.

Das geschieht am besten mithilfe eines **Tabellenkalkulationsprogramms**, wie **z.B. Excel**. Dazu werden die als txt- oder rtf-Datei exportierten Ergebnisse z.B. in Word geöffnet und dann manuell oder automatisiert[2] in die Tabelle übertragen. In den Export-Dateien folgen je nach Einstellung den Metaangaben zur Suche (Datum, Suchanfrage etc.) die Belege in KWIC-Ansicht und/oder in Volltext-Ansicht (vgl. Abb. 10). Die KWIC- bzw. Volltextbelege werden in die Excel-Tabelle übertragen, wobei **jeder Beleg eine Zeile** darstellt. Den ersten Spalten, in die die Referenznummer (siehe Beispiele (5) und (6) in Beispielbox oben) und der konkrete Beleg übertragen wird, folgen die Annotationsspalten, die sich aus der zu Beginn durchgeführten Operationalisierung ergeben.

ⓘ Das wichtigste bei der Annotation ist die **Einheitlichkeit**, für die eine gute und präzise Operationalisierung der Annotationskriterien nötig ist! Dazu sollten die Kriterien bzw. deren Operationalisierung als **Annotationsrichtlinien** formuliert werden, die später auch in der wissenschaftlichen Arbeit genau so erläutert werden.

Beispiel 1: *flink* und *flott*

Neben dem Beleg wird das zu klassifizierende Nomen noch einmal wiederholt, gefolgt vom Nomen-Typ (z.B. Objekte, Ereignisse, Zustände, Tropen) und ggf. Bemerkungen [Anmerkung: die gezeigten Annotationen sind für Demonstrationszwecke ausgedacht!]. Da die Untersuchung die beiden Adjektive *flink* und *flott* miteinander vergleichen soll, empfiehlt es sich, die Ergebnisse beider separat gestellter Suchanfragen in einer Tabelle zu annotieren. Dazu wird die Spalte „Adjektiv“ benötigt.

Referenz	Beleg	Adjektiv	Nomen	Typ	Bemerkungen
A10/JAN.02619 St. Galler Tagblatt, 14.01.2010, S. 44; Kein Heimsieg für UH Appenzell	Die heimische Verteidigung war oft einen Schritt langsamer als die **flinken Flügelspieler** aus Mörschwil.	flink	Flügelspieler	Objekt (belebt)	
A10/JAN.03699 St. Galler Tagblatt, 16.01.2010, S. 51; Alltag	Herzchen und lachende «Smilies» schmieren sie mit **flinken Fingern** auf Heckscheiben und Autotüren.	flink	Finger	Objekt (unbelebt)	
A10/JAN.03939 St. Galler Tagblatt, 18.01.2010, S. 38; Im Kulturforum den Affen lausen	Statt mit Lamento tut sie dies mit Witz und Charme, mit **flinken Pianistenfingern** und einer perfekten Chansonnièren-Stimme.	flink	Pianistenfinger	Objekt (unbelebt)	
A10/JAN.04505 St. Galler Tagblatt, 20.01.2010, S. 33; Junge Talente musizierten an der Kanti	Dorothea Federer malte mit **flinken Fingern** in «Manhatten Skyline» ein von Jazzrhythmen beeinflusstes Bild auf der Tastatur des Klaviers.	flink	Finger	Objekt (unbelebt)	
A10/JAN.07510 St. Galler Tagblatt, 30.01.2010, S. 11; Couture-Zauber	wie es die **flinken Näherinnen** im Pariser Atelier tun?	flink	Näherinnen	Objekt (belebt)	

[2] Das geht am einfachsten mit der von unserer Kollegin Anna Pryslopska entwickelten Online-Anwendung: https://anna-pryslopska.shinyapps.io/COSMAS2-Reader/

Bei der Annotation zeigt sich, **wie gut die Operationalisierung war**: Sind die Eigenschaften bzw. deren Ausprägungen klar definiert und abgegrenzt, geht die Annotation entsprechend schnell. Muss bei jedem Beleg dagegen lange überlegt werden, wie er annotiert wird, sollte die Operationalisierung angepasst werden. Generell gilt: je weniger mögliche Ausprägungen pro Annotationskriterium (= Eigenschaftsspalte) zur Auswahl stehen, desto einfacher die Zuordnung (und anschließend auch deren Auswertung).Natürlich lässt es sich nicht vermeiden, dass die Annotation in einzelnen Fällen schwerfällt. Wenn es sich um sehr **schwer zu interpretierende Treffer** handelt, deren Ungeeignetheit bei der manuellen Überprüfung vor Export übersehen wurde, dürfen diese auch an dieser Stelle noch aussortiert werden.

Weiterhin ist es möglich, dass sich im Laufe der Annotation herausstellt, dass die ursprünglich vorgesehenen Annotationskriterien um weitere oder andere ergänzt werden müssen.

> **Beispiel 2: *aber* und *nicht***
> Ursprünglich für die Untersuchung des Zusammenhangs von Kontrast und Negation vorgesehen war die Ermittlung der relativen Position von *nicht* zu *aber* („x, aber nicht y“ vs. „nicht x, aber y“) und des Skopus von *nicht* (Negation eines Attributs, eines Komplements, eines Adjunkts oder des Verbs) (siehe Abschnitt 1).
> Durch Belege wie in (7) bis (9) wird aber deutlich, dass es zusätzlich sinnvoll ist, den Umfang des Kontrastziels zu annotieren, also ob es sich um kontrastierte Sätze wie in (7), Phrasen wie in (8) oder potentiell elliptische Sätze wie in (9) handelt. Die Operationalisierung müsste also um die Eigenschaft „Umfang des Kontrastziels“ mit den Ausprägungen „vollständiger Satz“, „Phrase“ und „elliptischer Satz“ ergänzt werden.
> (7) Das Grauen lässt sich erahnen, **aber** begreifen lässt es sich **nicht**.
> (A97/APR.00483St. Galler Tagblatt, 25.04.1997)
> (8) «**nicht** naheliegend, **aber** dennoch denkbar»
> (A97/APR.00734 St. Galler Tagblatt, 26.04.1997)
> (9) Die Lage ist ernst, **aber nicht** hoffnungslos. → *aber [sie ist] nicht hoffnungslos*
> (A97/APR.00776 St. Galler Tagblatt, 26.04.1997)

Es empfiehlt sich, die eigenen **Annotationen** mit denen einer unabhängigen Person **abzugleichen**. Dazu kann eine kleine Auswahl der Belege mit den entsprechenden Operationalisierungen einer anderen Person gegeben werden und deren Ergebnis mit dem eigenen verglichen werden. Gibt es viele und große Abweichungen, sollte die Operationalisierung bzw. die Annotationskriterien angepasst werden. Bei größeren Studien empfiehlt es sich, diesen Prozess in Form eines umfangreicheren und aussagekräftigen **Inter-Annotator-Agreements** durchzuführen.

5. Die Auswertung

Nachdem sämtliche exportierte Belege annotiert wurden, wird die Annotation quantitativ ausgewertet. Dazu müssen die annotierten Eigenschaften bzw. deren Ausprägungen zunächst **quantifiziert** werden, also in Zahlen umgewandelt, die eine statistische Auswertung erlauben. 5.1. erläutert diesen Prozess anhand des Programms Excel. In 5.2. wird abschließend zusammengefasst, was beim Berichten der Korpusuntersuchung in der wissenschaftlichen Arbeit zu beachten ist.

5.1. Deskriptive Auswertung

Wie die Daten quantifiziert werden können, ist davon abhängig, von welchem **Datentyp** die annotierten Eigenschaften sind. Grundsätzlich sind zwei Datentypen zu unterscheiden: Kategorische Daten und metrische Daten.

Kategorische Daten:	Die Daten sind in Kategorien eingeteilt. • Nenndaten (englisch: *nominal data*): die Kategorien haben keine inhärente Ordnung (z.B. Geschlecht [m/w/n-b]) • Ordnungsdaten (*ordinal data*): die Kategorien haben eine inhärente, aber nicht-numerische Ordnung (z.B. Bildungsstand [Grundschule > weiterführende Schule > Studium]) Die Daten werden üblicherweise anhand der (absoluten oder relativen) **Häufigkeit** ausgewertet.
Metrische Daten: (*cardinal data*)	Die Daten sind numerische Angaben, die inhärent geordnet sind (z.B. Alter [in Zahlen], Wortlänge [Anzahl Buchstaben]). Die Daten werden üblicherweise anhand des **Mittelwerts** ausgewertet.

Da in einer Korpusuntersuchung meist mehrere Eigenschaften annotiert werden, ist es durchaus möglich, dass diese Eigenschaften unterschiedlichen Datentypen entsprechen.

Beispiel 1: *flink* und *flott*
Die zentralen Kriterien in dieser Untersuchung sind die Adjektive selbst, deren Ausprägung entweder zur Kategorie *flink* oder zur Kategorie *flott* gehört, und die Nomen-Typen, die sich in die Kategorien Objekte, Ereignisse, Zustände und Tropen einteilen lassen. Es handelt sich also um kategorische Daten, genauer gesagt Nenndaten.

Beispiel 2: *aber* und *nicht*
Auch in diesem Fall sind die untersuchten Eigenschaften (relative Position von *nicht* zu *aber*, Skopus von *nicht*, Umfang des Kontrastziels) vom Typ der kategorischen Daten. Da der Umfang des Kontrastziels inhärent geordnet ist (Phrase > elliptischer Satz > vollständiger Satz), aber nicht numerisch angegeben wird (es ist irrelevant, wie viele Wörter eine Phrase bzw. ein Satz enthält), handelt es sich hierbei um Ordnungsdaten.

Entsprechend der Datentypen findet nun die **deskriptive Auswertung** statt: Bei kategorischen Daten wird ausgezählt, wie häufig die einzelnen Kategorien innerhalb einer Annotationsspalte annotiert wurden. Bei metrischen Daten wird der Mittelwert (meist in Abhängigkeit zu anderen, kategorischen Daten) ermittelt.

Excel bietet hierfür das Mittel der **Pivot-Tabelle** („PivotTable“), die unter dem Reiter *Einfügen* auf ein neues Datenblatt eingefügt werden kann. Dabei handelt es sich um eine spezielle Tabellenform, in der Daten auf unterschiedliche Art dargestellt und ausgewertet werden können, ohne die Ausgangsdaten dabei verändern zu müssen. Dazu wird die Annotationstabelle als Quelle ausgewählt, die vorher allerdings als solche formatiert werden muss (inklusive Überschriften). Werden die Daten in der Quelltabelle im Nachhinein verändert (z.B. die Bezeichnungen verändert), wird dies in der Pivot-Tabelle automatisch angepasst.

Abb. 11 zeigt die **PivotTable-Felder**, mit deren Hilfe die Auswertung der Untersuchungsfrage entsprechend gestaltet werden kann. Im Bereich ***Feldname*** sind die Spalten (Annotationskriterien) aus der ursprünglichen Annotationstabelle aufgelistet. Diese können nun ausgewählt und den ***Spalten*** und ***Zeilen*** zugeordnet werden. Hierbei ist es üblich, die abhängige Variable (also die Eigenschaft, die beobachtet werden soll) den Spalten zuzuordnen und die unabhängige Variable (also die Eigenschaft, die die zu beobachtende Eigenschaft beeinflusst) den Zeilen.

Im Bereich ***Werte*** wird, in Abhängigkeit vom Datentyp, per Klick auf ⓘ festgelegt, wie die Eigenschaften ausgewertet werden sollen. Für kategorische Daten wird die Anzahl (Einstellung: „Zusammenfassen mit: *Anzahl*“) in absoluter (Einstellung: „Daten zeigen als: o*hne Berechnung*“) oder relativer (Einstellung: „Daten zeigen als: *%*“) Häufigkeit angegeben.

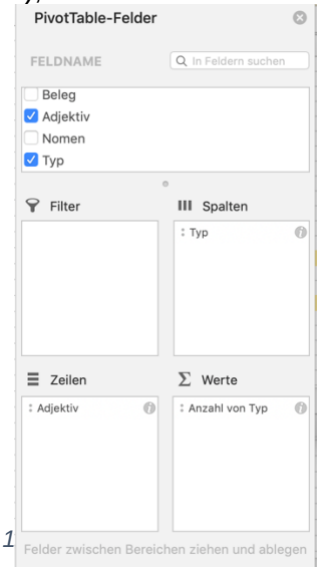


Abb. 1

Für metrische Daten wird üblicherweise der Mittelwert angegeben (Einstellung: „Daten zeigen als: *Mittelwert*“).

Die Auswertung kann mehrere Annotationskriterien (also Spalten in der ursprünglichen Annotationstabelle) auf einmal erfassen und kombinieren. Sollen viele Annotationskriterien auf einmal betrachtet werden, kann der Bereich ***Filter*** nützlich werden.

Analog zur Pivot-Tabelle lässt sich die Verteilung der Daten darüber hinaus mit dem **PivotChart** auch **grafisch** darstellen (siehe Beispielkasten unten).

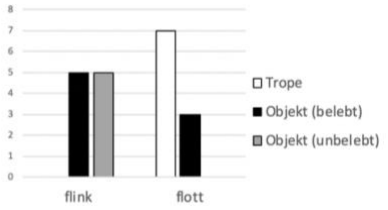
Die Verteilung der annotierten Kriterien innerhalb der Korpusdaten mit solchen deskriptiven Mitteln muss nun bezüglich der Fragestellung **inhaltlich ausgewertet** werden: Was bedeutet die Verteilung der Daten in Hinblick auf die Fragestellung? Liefern die Daten Evidenz für die Hypothese oder nicht? Und was bedeutet das für die linguistische Theorie?

Beispiel 1: *flink* und *flott*

Ausgewertet werden soll, wie häufig welcher Nomen-Typ in Abhängigkeit zum Adjektiv annotiert wurde. Der Feldname *Nomen-Typ* wird also als abhängige Variable den Spalten, *Adjektiv* als unabhängige Variable den Zeilen zugeordnet (siehe auch Abb. 11). Da es sich um kategorische Daten handelt, soll der Wert die Häufigkeit der Nomen-Typen je Adjektiv zeigen.

Die Auswertung der Annotation ergibt dabei, dass jeweils lediglich zwei der Nomen-Typen annotiert wurden: belebte und unbelebte Objekte bei *flink*, bzw. belebte Objekte und Tropen bei flott.

Anzahl von Typ	Spaltenbeschriftungen			
Zeilenbeschriftungen	Objekt (unbelebt)	Objekt (unbelebt)	Trope	Gesamtergebnis
flink	5	5	/	10
flott	/	3	7	10
Gesamtergebnis	**5**	**8**	**7**	**20**



Wären diese Ergebnisse echt und umfangreicher, ließe sich also für die Untersuchungsfrage Folgendes schlussfolgern: Die beiden Adjektive *flink* und *flott* weisen einen Bedeutungsunterschied auf. Während *flink* gleichermaßen belebte und unbelebte Objekte modifiziert, modifiziert *flott* vornehmlich Tropen und belebte Objekte.

ⓘ Die hier besprochene **deskriptive Auswertung** der Ergebnisse lässt lediglich Aussagen über eine kleine Menge an Daten zu (nämlich die Anzahl an exportierten und annotierten Belegen). Um zu untersuchen, ob diese Ergebnisse für das Sprachsystem bzw. dessen untersuchten Ausschnitt aussagekräftig, also *statistisch signifikant*, sind, ist zusätzlich eine **inferenzstatistische Auswertung** nötig, wie sie beispielsweise in den Kapiteln 5 und 6 in Stefanowitsch (2020) ausführlich und anschaulich erläutert wird (siehe weiterführende Literatur).

5.2. Darstellung der Korpusuntersuchung in der wissenschaftlichen Arbeit

In einem letzten Schritt werden die Ergebnisse der Korpusuntersuchung schließlich in der wissenschaftlichen Arbeit dargestellt und erläutert.

Innerhalb der wissenschaftlichen Arbeit entspricht die Korpusuntersuchung üblicherweise einem separaten Kapitel. Dieses ist entsprechend der Arbeitsschritte (vgl. Abb. 1) folgendermaßen strukturiert:

X.1. Fragestellung und Vorgehen

- Wiederholung der Fragestellung bzw. deren operationalisierte Version

- Korpuswahl benennen (Archiv und Korpus/Korpora) und begründen
- Suchanfrage(n) angeben, ggf. Prozess der Optimierung erläutern (vgl. Kapitel 3)

X.2. Annotationskriterien

- Beschreibung der verwendeten Annotationsrichtlinien: Eigenschaften und deren Ausprägungen werden möglichst anhand konkreter Korpusbelege demonstriert

X.3. Ergebnisse

- Angabe, wie viele Treffer die Suche ergab und wie viele davon in die Analyse einbezogen werden – manuell aussortierte Treffer (siehe 4.1.) werden (zusammengefasst) gezeigt und begründet
- Deskriptive (und ggf. inferenzstatistische) Auswertung: die Daten werden wertungsfrei und objektiv beschrieben und möglichst mithilfe von Tabellen und Abbildungen veranschaulicht

X.4. Diskussion

- Erläuterung besonders repräsentativer oder interessanter Belege
- Abgleich der Ergebnisse mit Fragestellung und Hypothese(n): Was konnte (nicht) gezeigt werden?

ⓘ Um Korpusbelege in der wissenschaftlichen Arbeit zu verwenden, muss immer die **Referenznummer** angegeben werden, vgl. Beispiel (9) aus Abschnitt 4.2.:
(9) Die Lage ist ernst, aber nicht hoffnungslos.
(A97/APR.00776 St. Galler Tagblatt, 26.04.1997)
Zudem müssen die **Korpora als Quellen** angegeben werden. Die zitierfähige Angabe der Korpora des IDS lautet:
Das Deutsche Referenzkorpus DeReKo, http://www.ids-mannheim.de/kl/projekte/korpora/, am Institut für Deutsche Sprache, Mannheim.
Die zitierfähige Angabe der Software COSMAS II lautet:
COSMAS I/II (Corpus Search, Management and Analysis System), http://www.ids-mannheim.de/cosmas2/, © 1991-2016 Institut für Deutsche Sprache, Mannheim.

6. Nützliche Links und weiterführende Literatur

Link zum Portal COSMAS II_{web}: http://www.ids-mannheim.de/cosmas2/

Weitere Hinweise zur Anfragesyntax: http://www.ids-mannheim.de/cosmas2/web-app/hilfe/suchanfrage/eingabe-zeile/syntax/

Übersicht über morphosyntaktische Annotationen in den Tagger-Sets:

- stts: http://www.ids-mannheim.de/cosmas2/projekt/referenz/stts/morph.html
- CONNEXOR: http://www.ids-mannheim.de/cosmas2/projekt/referenz/connexor/morph.html

Link zur Online-Anwendung: Umwandlung txt-Exportdatei in Excel-Tabelle

https://anna-pryslopska.shinyapps.io/COSMAS2-Reader/

Einführungen in die Korpuslinguistik:

LEMNITZER, Lothar & Heike ZINSMEISTER (2006): *Korpuslinguistik. Eine Einführung*. Tübingen: Gunter Narr Verlag.

SCHERER, Carmen (2006). *Korpuslinguistik*. Heidelberg: Universitätsverlag Winter.

STEFANOWITSCH, Anatol (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.