EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät

Deutsches Seminar

Regina Zieleke

Handreichung zur Korpusrecherche

in den DWDS-Korpora

Die Korpora des DWDS (Digitales Wörterbuch der deutschen Sprache) enthalten über 28 Milliarden Wörter in historischen und gegenwartssprachlichen Korpora. Neben schriftlichen Daten stehen hier auch (verschriftlichte) mündliche Daten zur Verfügung. Darüber hinaus sind alle Korpora mit Metadaten (z.B. Textsorte, Autor, Erscheinungsdatum) versehen und die darin enthaltenen Wörter sind lemmatisiert und mit Wortartinformationen versehen.

Die elektronische Recherche ist über die Internetseite http://dwds.de/r mit der linguistischen Suchmaschine DDC zugänglich. Die Verwendung der meisten über DWDS verfügbaren Korpora erfordert keine Registrierung. Es gibt jedoch Einschränkungen, z.B. in der erlaubten Komplexität der Suchanfrage. Der Zugang zu Spezialkorpora und komplexere Suchanfragen sind nach kostenfreier Registrierung möglich, bei der Sie angeben, dass Sie die Daten zu Forschungszwecken verwenden möchten.

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 Forschungsliteratur 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 am besten ermöglicht. Hier kann berücksichtigt werden, oder ob bestimmte zeitliche Aspekte (diachrone Untersuchung) oder textsortenspezifische Eigenschaften für die Suche relevant sind, und welches Medium untersucht werden soll (mündlich, schriftlich oder beides).

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.

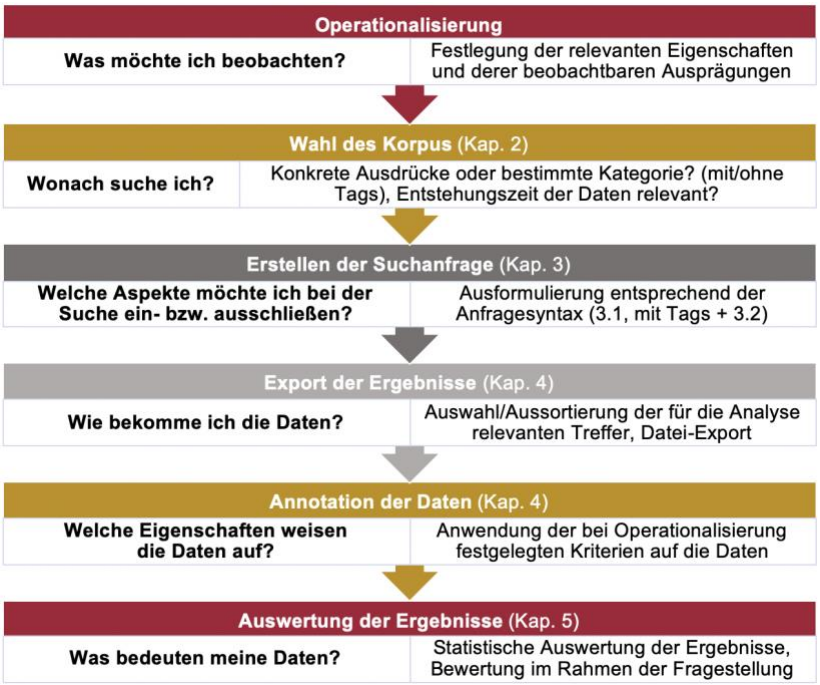


Abb. 1 Übersicht: Ablauf einer Korpusuntersuchung

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

Das DWDS ist eine Sammlung von Korpora, über die insgesamt mehr als 28 Milliarden Tokens in verschiedenen Korpora zugänglich sind (über die Website: https://www.dwds.de/r). Sämtliche Tokens in allen Korpora sind „getaggt“, also zu jedem Wort sind morphosyntaktische Informationen enthalten. Die Unterschiede zwischen den verfügbaren Korpora betreffen die Textsorten, ihre Geltungsbereiche (Referenz- oder Spezialkorpora), das Sprachmedium (schriftlich oder mündlich) und das Sprachstadium (historisch oder Gegenwartssprache).

Abb. 2 Suchfeld im DWDS

Referenzkorpora

Referenzkorpora sind zeitlich und hinsichtlich der Textsortenverteilung ausgewogen und gelten somit als repräsentativ für eine Sprache in ihrer Gesamtheit. Zu ihnen zählen die beiden **DWDS-Kernkorpora** je für das 20. und das (laufende) 21. Jahrhundert sowie das **Deutsche Textarchiv (DTA)**.

Die DWDS-Kernkorpora umfassen die vier Textsorten Belletristik, Gebrauchsliteratur, Wissenschaft und Zeitung aus dem jeweiligen Jahrhundert. Wie in Abb. 2 zu sehen ist, können einzelne Textsorten ausgewählt bzw. ausgeschlossen werden. Da die DWDS-Kernkorpora pro Jahrzehnt erstellt werden, sind Texte im (deutlich kleineren) DWDS-Kernkorpus 21 nur bis einschließlich 2010 verfügbar.

Die Texte im DTA sind ebenfalls ausgewogen, aber mit einem Schwerpunkt ab dem frühen 16. bis zum frühen 20. Jahrhundert. Sie eignen sich also für diachrone Betrachtungen.

Zeitungskorpora

Die über DWDS verfügbaren Zeitungskorpora enthalten retrodigitalisierte oder rein digital erstellte Texte großer Tages- und Wochenzeitungen, darunter **Berliner Zeitung**, **der Tagesspiegel** und **die ZEIT**. Zwar enthalten auch die DWDS-Kernkorpora Zeitungstexte, jedoch sind dort aufgrund der angestrebten Repräsentativität weniger Zeitungstexte vorhanden. Darüber hinaus sind einige Zeitungskorpora aus aktuelleren Zeiträumen; das ZEIT-Korpus beispielsweise ist bis zum Jahr 2018 durchsuchbar. Die Zeitungskorpora sind die umfangreichsten Korpora (z.B. der Tagesspiegel mit über 520 Millionen Tokens, verglichen mit ca. 121,5 Millionen Tokens im DWDS-Kernkorpus).

Spezialkorpora

Die Spezialkorpora sind für bestimmte Sprachbereiche zusammengestellt. Ohne Anmeldung zugänglich sind dabei folgende:

- **DTA-Erweiterungen** (weitere Texte aus dem Zeitraum von 1465 bis 1969, die aufgrund der angestrebten Ausgewogenheit nicht Teil des DTA sind, z.B. Sammlungen von Humboldt-Schriften oder historische Zeitschriften)
- **Archiv der Gegenwart** (Zeitschrift zu tagespolitischen Ereignissen aus Deutschland und der Welt von 1931 bis 2000)
- **Polytechnisches Journal** (digitalisierte Version aller 375 Bände der 1820 bis 1931 erschienenen Fachzeitschrift)
- **Filmuntertitel** (große Sammlung von Film- und Serienuntertiteln auf Basis des deutschsprachigen Teils der Communityplattform opensubtitles.org)
- **Gesprochene Sprache** (Transkripte von Reden, Parlamentsprotokollen und Interviews aus dem 20. Jahrhundert)
- **DDR** (ca. 1 100 Texte von 1949 bis 1990 aus der DDR; u.a. Reden, Rundfunkansprachen)
- **Politische Reden** (Reden bedeutender politischer Persönlichkeiten von 1982 bis 2020)

Für Untersuchungen des gesprochenen Deutschs eignet sich neben dem Korpus Gesprochene Sprache eingeschränkt auch das umfassendere Filmuntertitel-Korpus. Zu beachten ist dabei allerdings, dass die Dialoge gescriptet sind und damit vergleichbar mit Dialogen aus der Belletristik. Je nach Untersuchungsthema und -umfang empfiehlt sich evtl. ein Ausweichen auf eine andere Korpussammlung, z.B. die Datenbank für Gesprochenes Deutsch (DGD), in der z.B. auch Dialektdaten zugänglich sind.

Webkorpora

Die Webkorpora basieren auf einer Auswahl von Webseiten aus Deutschland, Österreich und der Schweiz. Der Großteil der Webkorpora, darunter das **Webkorpus Ballsportarten** und das im Aufbau befindliche **Corona-Korpus**, ist erst nach Anmeldung zugänglich. Frei zugänglich ist aber das **Blogs-Korpus**, das aus auf verschiedenen Blogs veröffentlichten Beiträgen und Kommentaren besteht. Sinnvoll ist die Verwendung dieser Korpora für Forschungsziele bezüglich internetbasierter Kommunikation oder Sprachvariation.

Metakorpora

Metakorpora bündeln Daten aus verschiedenen Korpora mit ähnlichen Metadaten, sodass eine Recherche über ein einzelnes Korpus hinaus möglich ist.

Die Option **Referenz- und Zeitungskorpora** vereint die beiden Kernkorpora, das Deutsche Textarchiv sowie die drei Zeitungskorpora. Damit sind in ihnen ca. 1,1 Milliarden Wortformen enthalten, allerdings zum Preis starker Unausgewogenheit.

Die **Historischen Korpora** vereinen das DTA und dessen Erweiterungskorpus mit weiteren, sonst nur nach Anmeldung zugänglichen historischen Korpora, wie bspw. den Schriften von Alexander von Humboldt und digitalisierte Texte historischer Zeitschriften (neben dem Polytechnischen Journal auch Die Grenzboten). Hier sind ebenfalls knapp 1 Milliarde Wortformen enthalten.

In allen Korpora kann der **Untersuchungszeitraum eingeschränkt** werden. Bei einigen Korpora lassen sich weitere Einschränkungen festlegen, so der Ausschluss von einzelnen Textsorten bei den Referenzkorpora, einzelne Teilkorpora bei den Metakorpora und bestimmte Sammlungen bei den DTA-Erweiterungen.

Für die meisten Untersuchungsfragen eignen sich die Referenz- und Zeitungskorpora, die einen repräsentativen Überblick an Sprachdaten zu den meisten Phänomenen des geschriebenen Deutschs liefern. Die Verwendung von Spezialkorpora bedarf immer einer Begründung, die sich an der Untersuchungsfrage orientiert.

Beispiel 1: *flink* und *flott*
Erwartbar ist, dass die Adjektive *flink* und *flott* häufiger im geschriebenen Deutsch verwendet werden, aber vielleicht nicht unbedingt in Blogs. Geeignet scheinen also insbesondere die DWDS-Kernkorpora und Zeitungskorpora. Da die Untersuchungsfrage keine diachrone Entwicklung etwaiger Bedeutungsunterschiede, sondern den „Ist-Zustand" betrifft, sollte das DTA oder das Metakorpus Referenz- und Zeitungskorpora nur für einen bestimmten Zeitraum durchsucht werden, etwa für die letzten 100 Jahre.

Beispiel 2: *aber* und *nicht*
Für die Untersuchung des Zusammenhangs von Kontrast und Negation ist keine konkrete Einschränkung in der Korpuswahl nötig. Referenz- und Zeitungskorpora eignen sich ebenso wie Spezialkorpora zum gesprochenen Deutsch oder politische Reden.

3. Die Suchanfrage

Nach Festlegung des Korpus/der Korpora wird im Eingabefeld nun die Suchanfrage entsprechend der Anfragesyntax formuliert. Die Suchanfrage sollte **so präzise wie möglich** formuliert sein, damit die manuelle Überprüfung und Aussortierung ungeeigneter Treffer im Nachhinein minimiert wird.

ⓘ 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?

Die Suchanfrage kann bereits aus einem **konkreten Ausdruck** bestehen, z.B. *Universität*. Bei einer Suche nach mehreren Ausdrücken, z.B. *Universität Tübingen*, wird per Default eine Phrasensuche angenommen, die die Ausdrücke unmittelbar aufeinanderfolgend sucht, vgl. Abb. 3.

…

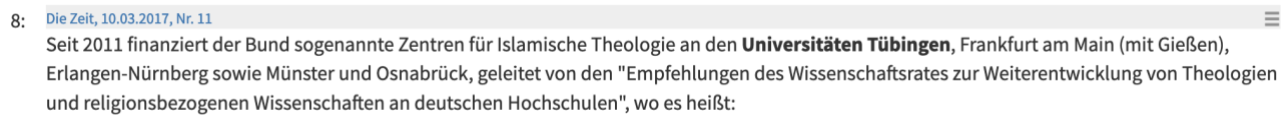
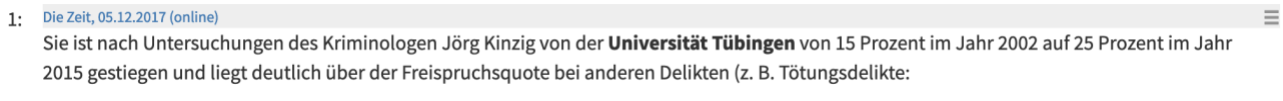


Abb. 3 Einfache Suchanfrage

Für die **Anzeige der Treffer** stehen drei Optionen zur Verfügung: KWIC, voll und maximal (siehe Auswahlfeld unter Korpuswahl). In der KWIC-Ansicht („key word in context“) werden die gesuchten Ausdrücke in ihrem unmittelbaren Kontext angezeigt und untereinander ausgerichtet. Somit lässt sich leichter ein Überblick über verschiedene Wortformen erhalten. Die Volltext-Ansicht zeigt die gesuchten Ausdrücke innerhalb eines vollständigen Satzes, wohingegen die maximal-Ansicht einen größeren Kontext mit je einem Vorgänger- und einem Folgesatz anzeigt.

Ansicht: KWIC

Ansicht: voll



Ansicht: maximal

Abb. 4 Trefferansichten: KWIC, voll und maxmal

Wie Treffer 8 in Abb. 3 zeigt, wird bei einer solchen „einfachen“ Suche nicht nur die konkret angegebene Wortform gesucht, sondern das **Lemma**. Neben der Grundform werden also auch flektierte Wortformen gesucht, so *Universitäten Tübingen*. Die Suchanfrage über DWDS kann aber noch viel mehr. Dazu wird eine **Anfragesyntax** verwendet, deren gebräuchlichste Operatoren in 3.1. in ihren Grundzügen erläutert werden. In 3.2. werden die im DWDS verfügbaren morphosyntaktischen Tags erläutert.

3.1. Allgemeine Anfragesyntax: Operatoren

Für die komplexe Suche mit mehreren Suchbegriffen stehen eine Reihe von **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 über die Online-Hilfe abgerufen werden.

Wortformoperatoren

Wie bereits erwähnt ist die Standardeinstellung bei der Suche in den DWDS-Korpora die Lemma-Suche, d.h. es wird nach allen Flexionsformen eines Ausdrucks gesucht. Groß- und Kleinschreibung wird dabei nicht berücksichtigt. Zur Suche nach einer konkreten Wortform wird der

Wortformoperator `@` vor das Wort gegeben. In diesem Fall wird die exakte Zeichenkette gesucht, die dem Operator folgt, inklusive Groß- und Kleinschreibung.

> **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`.

Ein weiterer Operator, der die Wortform des Suchbegriffs betrifft, ist der **Platzhalteroperator `*`.** Dieser ist ein Platzhalter für eine beliebige Menge von Zeichen. Er ist nicht in der Position im gesuchten Ausdruck oder auf eine bestimmte Wortart beschränkt. Auf diese Art kann somit nach Flektions- ebenso wie nach Kompositions- und Derivationsformen gesucht werden, siehe Beispiele:

Was?	Wozu?	Beispiel
`@`	Konkrete Wortform	`@Tische` → ~~*Tisch,*~~ *Tische*~~*, Tisches, Tischtennis*~~
`*`	Platzhalter für 0 bis n Zeichen	`Tisch*` → *Tisch, Tische, Tisches, Tischtennis, …*
		`*tisch` → *Schreibtisch, Nebentisch, theoretisch, …*
		`ver*` → *versucht, verankert, verstehen, …*
		`*ung` → *Romanfassung, Achtung, Nebenwirkung,* ~~*Romanfassungen, Nebenwirkungen*~~ *…*

Bei der Suche mit dem Platzhalteroperator ist zu beachten, dass der Operator lediglich eine Zeichenmenge betrifft, nicht aber linguistische Eigenschaften (dazu siehe morphosyntaktische Tags in Abschnitt 3.2). Bei der Suche nach Komposita tauchen somit eben nicht nur Komposita auf, sondern alle Tokens, die diese Zeichenkette beinhalten (vgl. *theoretisch* bei der Suche nach `*tisch`). Darüber hinaus verhält sich der Platzhalteroperator insofern dem Wortformoperator gleich, als dass nun Groß- und Kleinschreibung und konkrete Zeichenabfolge berücksichtigt werden. Die Suche nach *-ung*-Derivaten mittels `*ung` wird also ausschließlich Singularformen anzeigen.

> ⓘ Bei trennbaren Verben (Partikelverben) findet DWDS – sowohl mit als auch ohne Wortformoperator – **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 explizit separat gesucht werden. Da die DWDS-Suchmaschine jedoch unmittelbar aufeinander folgende Suchausdrücke als Phrasen versteht, werden mit dem Suchbefehl `kommen an` Treffer ausgeschlossen, bei denen zwischen den Verbteilen etwas steht (also *kommt an, kam an, kommen an*, aber ~~*kommt heute an, kommen mit dem Auto an*~~). Möchte man letztere Art von Treffern einschließen, muss ein Abstandsoperator verwendet werden (z.B. `kommen #>0 an`). Aber Vorsicht: Ohne weitere morphosyntaktische Informationen (siehe 3.2) unterscheidet DWDS an dieser Stelle nicht, ob es sich bei *an* um eine Verbpartikel oder eine Präposition handelt!

Abstandsoperatoren

Abstandsoperatoren ermöglichen die Suche nach mehreren Begriffen, die in einem bestimmten minimalen, maximalen oder exakten Abstand zueinander vorkommen. Dazu wird der **Abstandsoperator #** mit dem entsprechenden mathematischen Zeichen und einer Zahl für die Abstandsmenge kombiniert, vgl. nachfolgende Übersichtstabelle.

Was?	Wozu?	Beispiel
`<X> #=N <Y>`	Exakt N Ausdrücke Abstand zwischen X und Y	`@ein #=1 @Haus` → ~~*ein Haus*~~*, ein eigenes Haus, ein weiteres Haus,* ~~*ein weiteres eigenes Haus*~~
`<X> #>N <Y>`	Mindestens N Ausdrücke Abstand zwischen X und Y	`@ein #>1 @Haus` → ~~*ein Haus*~~*, ein eigenes Haus, ein weiteres eigenes Haus, ein Demokrat im Weißen Haus, …*
`<X> #N <Y>`	Maximal N Ausdrücke Abstand zwischen X und Y	`@ein #1 @Haus` → *ein Haus, ein eigenes Haus,* ~~*ein weiteres eigenes Haus, ein Demokrat im Weißen Haus*~~*, …*

Die Abstandssuche bezieht sich auf die Anzahl der Tokens. D.h. neben (linguistisch definierten) Ausdrücken werden auch Sonderzeichen, wie bspw. Kommata einbezogen. So ergibt die Suchanfrage `Universität #>1 Tübingen` neben Treffern wie *Es gibt eine Studie der Universitäten Linz und Tübingen* (Die Zeit, 29.10.2017, Nr. 44) auch *Tübingen hat keine Universität, Tübingen ist eine Universität* (Die Zeit, 06.09.2016 (online)).

> ⓘ Wie bereits erwähnt ist der Default für direkt aufeinanderfolgende Suchausdrücke eine Phrasenannahme, also ein Abstand von 1. Ein Leerzeichen in der Suchanfrage wird also als **Default-Abstandsoperator `#1`** interpretiert. Da die Anfragesyntax bei DWDS jedoch nur eine bestimmte Anzahl an Operatoren zulässt, kann es in Kombination mit anderen (z.B. logischen) Operatoren sinnvoll sein, die **Phrasensuche mit Anführungszeichen** (`"<Phrase>"`) anzuzeigen (somit ist für die Anfrage weniger Rechenaufwand nötig).

Auch eine Abstandssuche mit der Anzahl Null ist möglich und teilweise auch erforderlich: #0 kann dazu verwendet werden, einem Ausdruck morphosyntaktische Eigenschaften zuzuordnen (siehe 3.2).

Sonderzeichen

Einige Zeichen haben innerhalb der DWDS-Abfragesprache eine besondere Bedeutung. Um nach solchen Zeichen zu suchen, muss ihnen ein **Backslash \ vorangestellt** werden. Für eine Suche nach der Phrase *selten, fast nie* bspw. muss das Komma eigens in die Suchanfrage aufgenommen werden `selten \, fast nie` (die fehlerhafte Suche `selten fast nie` ergibt erwartungsgemäß keine Treffer).

> ⓘ Zeichen mit besonderer Bedeutung in der DWDS-Anfragesprache:
> & | # ^ ~ = $. ! ? , : ; @ % / \ ( ) { } [ ] < > * ' "

Da ein **Bindestrich** in der DWDS-Abfragesprache keine Sonderfunktion einnimmt, kann nach diesem ganz ohne Backslash gesucht werden. So ermöglicht die Anfrage `- und` bspw. eine Suche nach Koordinationen von Komposita mit gleichem Kopfwort, wie *Chemie- und Physiklehrer* oder *Schwellen- und Entwicklungsländern*.

Verknüpfungsoperatoren

Wie jede Computersprache verwendet auch die Anfragesyntax von DWDS **logische** Verknüpfungsoperatoren mit der Funktion **„und“ (`&&`), „oder“ (`||`) und „nicht“ (`&& !`)**. Entsprechend der Aussagenlogik können mehrere Verknüpfungsoperatoren mithilfe von Klammerungen miteinander kombiniert werden. Die Eingabe `Universität && !Tübingen` sucht z.B. nach dem Ausdruck *Universität* ohne *Tübingen* (also z.B. *Universität Stuttgart, Universität, Universität zu Köln*), die Anfrage `(Universität Tübingen) || (Universität Stuttgart)`

entsprechend *Universität Tübingen* oder *Universität Stuttgart* (oder beides, es handelt sich um eine inklusive Disjunktion). Aber Vorsicht: Die Suche `Universität (Tübingen || Stuttgart)` kreiert eine Fehlermeldung:

> ⓘ Die DWDS-Anfragesyntax **unterscheidet Verknüpfungen auf Anfrage- bzw. Tokenebene**:
> – Die logischen Operatoren **`&&`, `||` und `&& !`** können ausschließlich **auf Anfrageebene** verwendet werden. Es können zwar auch einzelne Ausdrücke gesucht werden, diese werden aber als voneinander unabhängige Anfragen („Klauseln “) behandelt, die nicht mehr durch weitere Operatoren ergänzt werden können. Dementsprechend bestehen zwischen den Anfragetermen beliebige Abstände innerhalb eines Satzes.
> – Für Verknüpfungen von Bedingungen **auf Tokenebene** müssen stattdessen die Operatoren **`with`, `withor` oder `without`** verwendet werden. Das konjunktive `with` hat dabei allerdings eine andere Funktion als das logische `&&` – es verknüpft nicht zwei Ausdrücke in beliebigen Abständen miteinander (dafür wird der Abstandsoperator `#>1` verwendet), sondern einen Ausdruck mit einer Eigenschaft (z.B. morphosyntaktische Tags, siehe 3.2).

Die Anfrage `(Universität Tübingen) || (Universität Stuttgart)` funktioniert, weil die umklammerten Ausdrücke als abgeschlossene Anfragen interpretiert werden. Das ändert sich, sobald weitere Operatoren hinzukommen (so auch im Falle einer Phrasensuche, bei der ja per Default ein Abstand von 1 zwischen den Tokens angenommen wird). Für die ökonomischere Suche nach nur einer Phrase mit den alternativen Ausdrücken *Tübingen* und *Stuttgart* muss die Anfrage also `Universität Tübingen withor Stuttgart` lauten.

Eine weitere Variante für solche Alternativfragen ist die Verwendung der **Alternativenklammern `{}`**. So erhält man das gleiche Ergebnis mit `Universität {Tübingen, Stuttgart}`. Diese Variante ist zudem mit weiteren Operatoren verknüpfbar, die auf die gesamte Alternativenmenge angewendet werden sollen. In der Suchanfrage `@die @{Stadt, Universität} Tübingen` bspw. hat der Wortformoperator `@` Skopus über beide Alternativen *Stadt* und *Universität*.

> ⓘ Die Möglichkeiten zur **Suche nach negierten Bedingungen** ist in der DWDS-Anfragesyntax nur **eingeschränkt** möglich. Die logische Negation ist nicht auf alle Anfragen anwendbar, wie oben erwähnt z.B. nicht auf Tokenebene, weshalb die Anfrage `die Stadt !Tübingen` eine Fehlermeldung erzeugt. Ein Ausweichen auf die Variante `without` ist wiederum auf Suchanfragen beschränkt, bei denen der auszuschließende Ausdruck folgt: `die Stadt without Tübingen` erzeugt Treffer der Form *die Stadt ~~Tübingen~~*. Möchte man dagegen Ausdrücke vor den Zielausdrücken ausschließen (z.B. *~~in~~ der Stadt*), ist `without` nicht die richtige Wahl: Die Anfrage `in without die Stadt` erzeugt genau gegenteilige Treffer (*in ~~der Stadt~~*). Trotzdem ist eine Suche nach *~~in~~ der Stadt* möglich, indem auf die Anfrageebene gewechselt wird: `"die Stadt" && !"in die Stadt"`.

Positionen im Satz

Alle DWDS-Korpora sind hinsichtlich von Metadaten indiziert, darunter Wortarten (siehe 3.2) und Satzposition. Der **Satzindex** in der DWDS-Abfragesyntax lautet **`$.=`**.

> ⓘ Ein **„Satz“** ist in DWDS 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**!

Innerhalb eines solchen „Satzes“ wird für jedes Token automatisch die Position gezählt: Das erste Wort im Satz hat die Position Null (`$.=0`), das darauffolgende die 1 etc.. Diese Position kann einem Ausdruck **mithilfe des Verknüpfungsoperators `with`** zugeordnet werden. Die Suchanfrage `ohne with $.=0` ergibt demnach Sätze, die mit *ohne* beginnen. Die Zählung kann auch rückwärts, also vom Satzende beginnend, vorgenommen werden. Dabei ist zu beachten, dass Satzzeichen wie oben erwähnt als eigenständige Tokens analysiert werden und in die Zählung aufgenommen werden. Das satzabschließende Zeichen trägt somit die Position `$.=-1`, das letzte Wort im Satz `$.=-2`. Entsprechend ergibt die Suchanfrage `nicht with $.=-2` Sätze, die mit *nicht* aufhören.

Beispiel 2: *aber* und *nicht*
Wie müssen wir die Suchanfrage formulieren, um den Zusammenhang von Kontrast und Negation zu untersuchen? Wir wissen nun:
– dass DWDS 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.
– Kommata zu den Sonderzeichen gehören, die mithilfe eines Backslashs gesucht werden müssen → `\, aber`
Aber und *nicht* sollten dann in freier Reihenfolge im selben „Satz“ auftreten. Wir wissen:
– dass die die logische Konjunktion `&&` auf Anfrageebene fungiert und automatisch eine beliebige Reihenfolge zwischen beiden Anfragen erzeugt.
– dass bei der Verwendung mehrerer Operatoren die Phrasensuche mit Anführungszeichen vereinfacht werden muss. Die Suchanfrage kann also lauten: → `"nicht" && "\, aber"`
Diese Suchanfrage ergibt im Korpus politische Reden (Zeitraum 1982 bis 2020) 11.381 Treffer. Durch die „automatisierte Satzannahme“ befinden sich darunter jedoch auch Treffer wie die folgenden:
(1) Wir wollen, dass am Ende das Tierwohl, der Respekt vor dem Tier, **aber** auch der Respekt vor dem Tierhalter und am Schluss der Respekt vor dem Verbraucher, der selbst entscheidet, was er verzehrt, zusammen gedacht werden und **nicht** die einzelnen Gruppen gegeneinander ausgespielt werden. (Rede von Julia Klöckner, 03.07.2020)
(2) Das ist heute ein besonderer Tag: **nicht** nur, weil wir über viel Geld beschließen, das wir aufnehmen und ausgeben, um dafür zu sorgen, dass die Konjunktur in diesem Land gut weiterentwickelt werden kann, sondern weil wir gleichzeitig am heutigen Tag eine weitere Entscheidung treffen, bei der ich mich sehr freue, dass sie zwar zufällig, **aber** vollständig richtig heute genauso mit ansteht. (Rede von Olaf Scholz, 02.07.2020
In (1) ist die Negation teil eines neuen, durch *und* eingeleiteten Satz, während sie in (2) nicht teil des *aber*-Korrelats, sondern des *sondern*-Korrelats ist. Solche Treffer können manuell aussortiert werden oder bereits in der Suchanfrage ausgeschlossen werden:
→ `"nicht" && "\, aber" && !"sondern" && !"und nicht"`
Diese Suchanfrage reduziert die Trefferzahl auf 10.199. (Da diese Suchanfrage mehr als 5 Operatoren enthält, sind hierfür Registrierung und Anmeldung nötig.)

3.2. Morphosyntaktische Tags

Sämtliche Tokens in den DWDS-Korpora sind maschinell mit sogenannten **Part-of-Speech-Tags** (PoS-Tags) versehen, also Informationen zu syntaktischen Kategorien, die etwas genauer ausdifferenziert sind als die üblicherweise unterschiedenen Wortarten. Die PoS-Tags sind nach dem **STTS-Tagset** (*Stuttgart-Tübingen-Tagset*) annotiert und über den **Index `$p=`** erfragbar.

Es kann allgemein nach Wörtern bestimmter Wortarten gesucht werden (z.B. Kardinalzahlen mit dem Suchbefehl `$p=CARD`) oder nach konkreten Ausdrücken, die die morphosyntaktischen Eigenschaften

aufweisen (z.B. der Ausdruck *der* als Relativpronomen). **Eigenschaften werden mit dem Verknüpfungsoperator `with` zugeordnet**, also z.B. `@der with $p=PRELS`. Selbstverständlich können auch die anderen Operatoren mit den PoS-Tags kombiniert werden. So erzeugt die Anfrage `$p=VAFIN with $.=0` bspw. Sätze, die mit finiten Hilfsverben beginnen (z.B. *Haben Sie vielen Dank und bleiben Sie gesund!* (Rede von Heiko Maas, 18.06.2020) oder *Ist Hölderlin noch aktuell?* (Rede von Frank-Walter Steinmeier, 23.05.2020)).

Darüber hinaus können die Tags selbst mit dem **Platzhalteroperator `*`** abgeändert werden, wenn die genaue Ausprägung nicht relevant ist. Bspw. ergibt die Suche `$p=P* $p=NN` Kombinationen aus (irgendwelchen) Pronomen und Nomen von *warum Menschen* bis *etwas Entscheidendes*.

Hier eine nach Wortarten sortierte **Übersicht** der ergfragbaren PoS-Tags in den DWDS-Korpora:

Verben: `$p=V*`

- Vollverben: `$p=VV*`
  - `VVFIN` finites Verb, voll *[du] gehst, [wir] kommen [an]*
  - `VVIMP` Imperativ, voll *komm [!]*
  - `VVINF` Infinitiv, voll *gehen, ankommen*
  - `VVIZU` Infinitiv mit „zu“, voll *anzukommen, loszulassen*
  - `VVPP` Partizip Perfekt, voll *gegangen, angekommen*
- Auxiliarverben: `$p=VA*`
  - `VAFIN` finites Verb, Auxiliar *[du] bist, [wir] werden*
  - `VAIMP` Imperativ, Auxiliar sei [ruhig!]
  - `VAINF` Infinitiv, Auxiliar werden, sein
  - `VAPP` Partizip Perfekt, Auxiliar gewesen
- Modalverben: `$p=VM*`
  - `VMFIN` finites Verb, modal *dürfen*
  - `VMINF` Infinitiv, modal *wollen*
  - `VMPP` Partizip Perfekt, modal *gekonnt, [er hat gehen] können*

Nomen: `$p=N*`

- `NN` normales Nomen *Tisch, Herr, [das] Reisen*
- `NE` Eigennamen *Hans, Hamburg, HSV*

Artikel: `$p= ART` (bestimmter oder unbestimmter Artikel *der, die, das, ein, eine*)

Adjektive: `$p=ADJ*`

- `ADJA` attributives Adjektiv *[das] große [Haus]*
- `ADJD` adverbiales oder prädikatives Adjektiv *[er fährt] schnell, [er ist] schnell*

Adverbien: `$p=ADV` (*schon, bald, doch*)

Präpositionen: `$p=APP*`

- `APPR` Präposition *in [der Stadt], ohne [mich]*
- `APPRART` Präposition mit Artikel *im [Haus], zur [Sache]*
- `APPO` Postposition *[ihm] zufolge, [der Sache] wegen*

Pronomen: `$p=P*`

- Personalpronomen: `$p=PPER` (*ich, er, ihm, mich, dir*)
- Reflexivpronomen: `$p=PRF` (*sich, einander, dich, mir*)
- Demonstrativpronomen: `$p=PD*`
  - `PDS` substituierendes Demonstrativpronomen *dieser, jener*
  - `PDAT` attribuierendes Demonstrativpronomen *jener [Mensch]*
- Indefinitpronomen: `$p=PI*`
  - `PIS` substituierendes Indefinitpronomen *keiner, viele, man, niemand*
  - `PIAT` attribuierend, ohne Determiner *kein [Mensch], irgendein [Glas]*
  - `PIDAT` attribuierend, mit Determiner *[ein] wenig [Wasser], [die] beiden [Brüder]*
- Possessivpronomen: `$p=PPOS*`
  - `PPOSS` substituierendes Possessivpronomen *meins, deiner*
  - `PPOSAT` attribuierendes Possessivpronomen *mein [Buch], deine [Mutter]*
- Relativpronomen: `$p=PREL*`
  - `PRELS` substituierendes Relativpronomen *[der Hund,] der*
  - `PRELAT` attribuierendes Relativpronomen *[der Mann,] dessen [Hund]*
- Interrogativpronomen: `$p=PW*`
  - `PWS` substituierendes Interrogativpronomen *wer, was*
  - `PWAT` attribuierendes Interrogativpronomen *welche [Farbe], wessen [Hut]*
  - `PWAV` adverbiales Interrogativ- oder Relativpronomen *warum, wo, wann, woher, wobei*
- Konnektivpronomen: `$p=PAV` (Pronominaladverbien *dafür, dabei, deswegen, trotzdem*)

Konjunktionen: `$p=KO*`

- `KOUI` unterordnend, mit „zu“ und Infinitiv *um [zu leben], anstatt [zu fragen]*
- `KOUS` unterordnend, mit Satz *weil, daß, damit, wenn, ob*
- `KON` nebenordnend *und, oder, aber*
- `KOKOM` vergleichend *als, wie*

Partikeln: `$p= PTK*`

- `PTKZU` „zu“ vor Infinitiv *zu [gehen]*
- `PTKNEG` Negationspartikel *nicht*
- `PTKVZ` abgetrennter Verbzusatz *[er kommt] an, [er geht] mit*
- `PTKANT` Antwortpartikel *ja, nein, danke, bitte*
- `PTKA` Partikel bei Adjektiv/Adverb *am [schönsten], zu [schnell]*

Beispiel 1: *flink* und *flott*
Wie muss die Suchanfrage nach der Kombination der Adjektive *flink* bzw. *flott* mit darauffolgendem Nomen aussehen? Wir wissen nun:
– dass die verschiedenen Flexionsformen der Adjektive automatisch gesucht werden.
– dass Alternativen auf Tokenebene entweder mit `withor` oder mit `{}` gesucht werden müssen → `flink withor flott` oder `{flink, flott}`
– dass Nomen mit dem PoS-Tag `$p=N*` gesucht werden. Da Eigennamen für die Untersuchungsfrage auszuschließen sind, sollte der Tag spezifiziert werden → `$p=NN`
Im Metakorpus Referenz- und Zeitungskorpora (Zeitraum 1918 bis 2018) ergibt `{flink, flott} $p=NN` 5779 Treffer. Für eine Untersuchung sollten die Suchanfragen getrennt werden und beispielsweise 200 Datensätze für je `flink $p=NN` (1694 Treffer) und `flott $p=NN` (4085 Treffer) annotiert werden.

4. Die Ergebnisse

Deutet die Anzeige der Treffer darauf hin, dass die Suchanfrage so präzise wie möglich gestellt ist, sodass – wenn überhaupt – nur wenige Treffer manuell aussortiert werden müssen, werden die Treffer exportiert, aufbereitet und annotiert.

4.1. Export und Aufbereitung der Ergebnisse

Um die Daten im Anschluss an die Suche weiter zu verwenden, müssen sie exportiert werden. Dazu können Trefferformat, Trefferanzahl, Datenformat und Art der Ausgabe festgelegt werden.

Abb. 5 Exportoptionen für die Ergebnisse

Entsprechend der Ergebnisanzeige kann für das **Trefferformat** zwischen den Optionen voll, KWIC und maximal gewählt werden. Der Umfang der Daten ist dabei für voll und KWIC identisch (ein „Satz“), lediglich die Anzeige ändert sich (im Fall von KWIC sind die Suchausdrücke („Hits“) in einer separaten Spalte vom restlichen Satz gelöst, was sich für unser Beispiel 1 (*flink/flott*) eignen würde).

Die **Trefferanzahl** hängt von der Untersuchungsfrage und -methode ab: bei qualitativen Arbeiten reichen je nach Phänomen für Hausarbeiten bereits 50 Datensätze. Für Arbeiten mit quantitativen Anteilen sollten mindestens 200 Datensätze analysiert werden (oder eben die maximale Trefferzahl bei kleineren Sprachphänomenen). Da unter Umständen einzelne Treffer manuell aussortiert werden müssen, sollten immer mehr Datensätze exportiert werden, als anvisiert.

Für das **Datenformat** empfiehlt sich eine CSV-Date als **Download**. CSV steht für *komma-getrennte Werte* (comma separated values) und ist in gängigen **Tabellenkalkulationsprogrammen**, wie **z.B. Excel** bearbeitbar. Dazu müssen die Daten jedoch **aufbereitet** werden:

Bei Export sind die Daten dem Dateiformat entsprechend durch Kommata getrennt und müssen über die Option *Daten > Text in Spalten* umformatiert werden, vgl. Abb. 6:

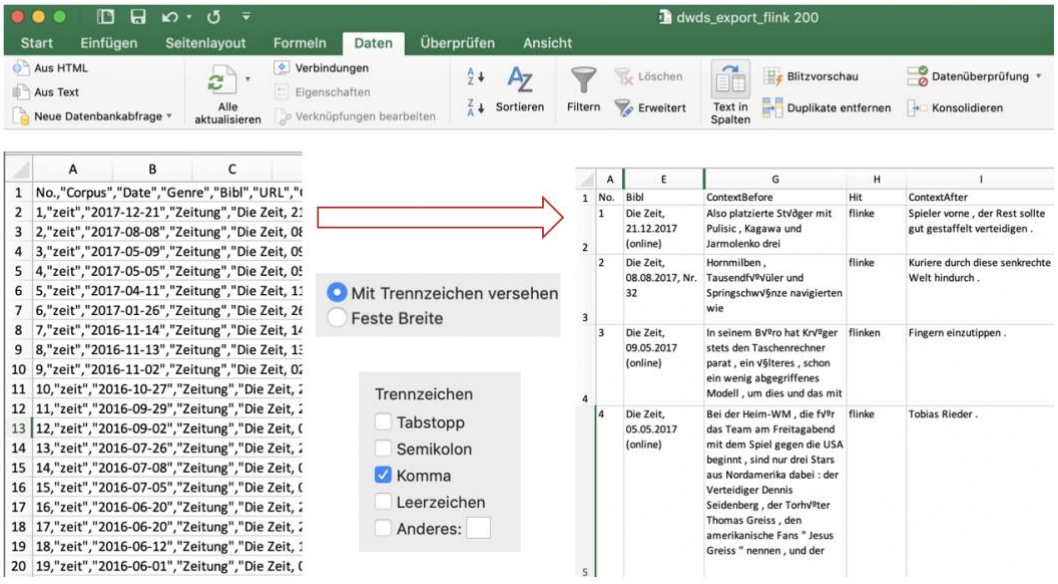


Abb. 6 Aufbereitung der Exportdatei in Excel

Von diesen Rohdaten sollten sämtliche Spalten beibehalten werden, insbesondere die **Angaben zur Herkunft des Treffers (Spalte „Bibl“)**. Diese Herkunftskürzel müssen für jeden Korpusbeleg, der in der Arbeit diskutiert wird, angegeben werden (vgl. (1) und (2) im Beispielkasten auf Seite 11; siehe 5.2).

4.2. Annotation der Ergebnisse

Nach dem Finden und Aufbereiten 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. Dazu werden in der aufbereiteten Exportdatei entsprechende **Annotationsspalten ergänzt**.

> ⓘ 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.

ContextBefore	Hit	ContextAfter	Adjektiv	Nomen	Typ	Anmerkungen
Also platzierte Stöger mit Pulisic , Kagawa und Jarmolenko drei	flinke	Spieler vorne , der Rest sollte gut gestaffelt verteidigen .	flink	Spieler	Animate	
Hornmilben , Tausendfüßler und Springschwänze navigierten wie	flinke	Kuriere durch diese senkrechte Welt hindurch .	flink	Kuriere	Animate	
In seinem Büro hat Krüger stets den Taschenrechner parat , ein älteres , schon ein wenig abgegriffenes Modell , um dies und das mit	flinken	Fingern einzutippen .	flink	Finger	Object	
Bei der Heim-WM , die für das Team am Freitagabend mit dem Spiel gegen die USA beginnt , sind nur drei Stars aus Nordamerika dabei : der Verteidiger Dennis Seidenberg , der Torhüter Thomas Greiss , den amerikanische Fans " Jesus Greiss " nennen , und der	flinke	Tobias Rieder .	flink	Tobias Rieder	Animate	Eigenname
Ein Anfangfünfziger mit	flinken	Augen , breiten Schultern und entschiedener Vitalität .	flink	Augen	Object	
Und auch das Programm des Eröffnungskonzerts stolpert ein wenig über seine	flinken	Füße , wenn es vor der Pause Britten , Dutilleux , Cavalieri , Zimmermann , Prätorius , Liebermann , Caccini , Messiaen nahtlos , atemlos aneinanderreiht und nach der Pause Wagner , Rihm , Beethoven (die Neunte !) ebenso .	flink	Füße	Object	

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.

> ⓘ 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 2: *aber* und *nicht*
Eine zusätzliche manuelle Aussortierung der Treffer ist hier aus (mindestens) zwei Gründen sinnvoll und nötig: Die in DWDS verwendete Definition von „Satz" ergibt auszusortierende Treffer wie in (3), bei dem die Negation in einem separaten, nicht durch *aber* verbundenen Satz auftaucht, und der Skopus der Negation kann höher sein als das Kontrastziel von *aber* wie in (4).

(3) Ein Virus macht an den Grenzen **nicht** halt, und deswegen gilt es, dass ganz Europa sich abstimmt: die Europäische Union, **aber** auch die Nachbarländer darüber hinaus. (Rede von Jens Spahn, 12.02.2020)

(4) Und wenn wir als Demokratien sie **nicht** leisten, dann werden es andere tun – zu einem Preis, der auf Dauer deutlich höher sein wird für uns, **aber** auch für andere. (Rede von Heiko Maas, 29.06.2020)

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 (5) bis (7) wird aber deutlich, dass es zusätzlich sinnvoll ist, den Umfang des Kontrastziels zu annotieren, also ob es sich um kontrastierte Sätze wie in (5), Phrasen wie in (6) oder potentiell elliptische Sätze wie in (7) handelt. Die Operationalisierung müsste also um die Eigenschaft „Umfang des Kontrastziels" mit den Ausprägungen „vollständiger Satz", „Phrase" und „elliptischer Satz" ergänzt werden.

(5) Wir wissen, was zu tun ist, **aber** wir tun es **nicht**. (Rede von Gerd Müller, 13.02.2020)

(6) Gelockert, **aber** noch **nicht** locker (Rede von Alain Berset, 13.06.2020)

(7) Unsere Wirtschaft braucht Sie, aber nicht nur die. → *aber nicht nur die [braucht sie]* (Rede von Frank-Walter Steinmeier, 23.06.2020)

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 bei der Darstellung 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.

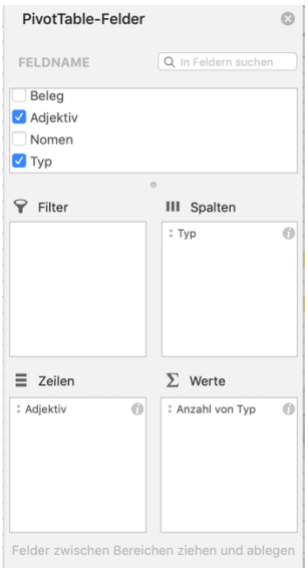


Abb. 8 zeigt die **PivotTable-Felder**, mit deren Hilfe die Auswertung der Untersuchungsfrage entsprechend gestaltet werden kann. Im Bereich ***Feldname*** sind die Spalten (Annotationskriterien) aus der ursprünglichen Annotationstabelle aufgelistet. Diese können nun ausgewählt und den ***Spalten*** 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. 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.

Abb. 7 PivotTable-Felder in Excel

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

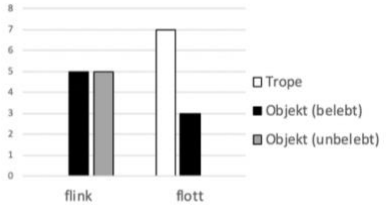
Die Verteilung der annotierten Kriterien innerhalb der Korpusdaten mit solchen deskriptiven Mitteln muss nun bezüglich der Fragestellung **inhaltlich ausgewertet** werden: Was bedeutet die Verteilung der Daten in Hinblick auf die Fragestellung? Liefern die Daten Evidenz für die Hypothese oder nicht? 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. 8). 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 und begründen (vgl. Kapitel 2)
- 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 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 das Kürzel für den **Quellennachweis** angegeben werden, das verwendete Korpus sowie das Zugriffsdatum:

(7) Unsere Wirtschaft braucht Sie, aber nicht nur die.
Rede von Frank-Walter Steinmeier, 23.06.2020, aus dem Korpus politische Reden (DWDS), abgerufen am <Datum>

Die zitierfähige Angabe des DWDS als Ganzes lautet:
DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunfts-system zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften, <https://www.dwds.de/>, abgerufen am <Datum>.

6. Nützliche Links und weiterführende Literatur

Link zur DWDS-Korpusabfrage: http://dwds.de/r

Weitere Hinweise zu einzelnen Korpora: https://www.dwds.de/d/korpora

Weitere Hinweise zur Anfragesyntax: https://www.dwds.de/d/korpussuche

Übersicht über morphosyntaktische Annotationen (Tagset): https://www.dwds.de/d/korpussuche#pos

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.