Arbeitsschritte: Vom Korpus über die Konkordanz zur Analyse

Dieses Tutorial stellt die einzelnen Arbeitsschritte vor, die bei einer „prototypischen" Korpusrecherche anstehen. Natürlich gibt es viele Spezialfälle, in denen wir diesen Schritten nicht 1:1 folgen können, sondern kreative und innovative Lösungen brauchen, die auf das jeweilige Problem zugeschnitten sind. Das kann beispielsweise dann der Fall sein, wenn es für unsere Fragestellung (noch) kein geeignetes Korpus gibt und wir selbst eines zusammenstellen müssen. Für die meisten Forschungsfragen, die im Rahmen von Seminar- und Abschlussarbeiten bearbeitet werden, reichen die existierenden Korpora aber völlig aus, und in aller Regel kann man sich dann grob an den folgenden Schritten orientieren.

1. Ein geeignetes Korpus wählen
Am Anfang jeder Korpusrecherche steht selbstverständlich eine Fragestellung. Ein Beispiel: Ich möchte untersuchen, in welchem Maße sprachkritische und sprachpuristische Bemühungen Einfluss auf den tatsächlichen Sprachgebrauch haben, und wähle als Beispiel die Varianten *Sinn machen* vs. *Sinn ergeben*, deren erstere von Sprachkritikern wie Bastian Sick bekanntlich gern gegeißelt wird, während die letztere propagiert wird (auch wenn sie genauso viel oder wenig Sinn macht wie die erste).
Nun gilt es, diese Fragestellung zu **operationalisieren**, d.h. einen konkreten Weg zu finden, sie mit Hilfe empirischer Methoden anzugehen. Eine Möglichkeit besteht darin, das Erscheinungsdatum von Bastian Sicks vielgelesenem Kolumnenbuch „Der Dativ ist dem Genitiv sein Tod", in dem auch *Sinn machen* vs. *Sinn ergeben* diskutiert wird, als „Ankerpunkt" zu wählen und konkret zu fragen, ob sich die Distribution der beiden Varianten im Sprachgebrauch nach dem Erscheinen des Buches verändert hat.
Welches Korpus ist geeignet, um diese Fragestellung zu untersuchen? Es muss zwei Kriterien erfüllen: Es muss einerseits ein gegenwartssprachliches Korpus sein, da ich mich ja für das heutige Deutsche interessiere; es muss andererseits aber auch ein diachrones (oder zumindest „mikro-diachrones") Korpus sein, das mir erlaubt, nicht nur Belege aus dem Jahr, sagen wir, 2017 zu betrachten, sondern vielmehr die Frequenzentwicklung der beiden Varianten um das Jahr 2000 herum zu verfolgen.
Hierfür bieten sich z.B. das Deutsche Referenzkorpus und das DWDS-Korpus an. Wir wählen für unsere Beispielanfrage das DWDS-Kernkorpus des 20. und 21. Jahrhunderts.

2. Eine Suchanfrage stellen
Im nächsten Schritt gilt es, die Belege zu finden, die wir finden möchten – und zwar möglichst *nur* diese Belege, damit sich die anschließende manuelle Korrekturarbeit in Grenzen hält. Wenn wir einzelne Wortformen suchen, etwa *daß* und *dass*, dann gestaltet sich die Suchanfrage extrem einfach und bringt nicht mehr Aufwand mit sich als eine einfache Google-Suche. Komplizierter wird es bei komplexeren Suchanfragen, etwa wenn wir alle Flexionsformen eines Wortes auf einmal suchen oder wenn wir gar nach syntaktischen Mustern suchen.
Bedenken Sie dabei: Anders als Google sind Korpus-Interfaces nicht „klug", sondern finden nur genau das, was man sucht - und das ist auch gut so, denn was bei einer Alltagssuche über Google sehr hilfreich und nützlich sein kann (also dass z.B. bei der Suche nach *Zeitung* auch so etwas wie *Magazin* gefunden wird und dass Tippfehler automatisch korrigiert werden), würde bei Korpusrecherchen zu zahlreichen Fehltreffern führen, die wir nicht wollen. Die Kehrseite des Aufwands, den Sie ggf. in das Erlernen der jeweiligen Suchabfragesyntax investieren müssen, ist also die Genauigkeit der Suchergebnisse - in der Regel wird genau das gefunden, was man sucht, und das ist extrem praktisch.
In der DWDS-Suchabfragesprache, in die man sich unter „Hilfe zur Suche" (https://www.dwds.de/d/suche, zuletzt abgerufen August 2017) einarbeiten kann, können wir

den *near*-Operator benutzen, um Belege zu finden, in denen das Lemma *Sinn* im Abstand von einer bestimmten Anzahl an Wörtern vom Lemma *machen* oder *ergeben* auftritt. Damit finden wir sowohl Sätze mit Verberst- oder Verbzweit- als auch mit Verbletztstellung, also sowohl *...weil das überhaupt keinen Sinn macht* als auch *Das macht überhaupt keinen Sinn* und *Macht das alles einen Sinn*? Die Abfrage mit *near*-Operator funktioniert wie folgt:

near(String1, String2, n)

wobei String1 und String2 für Einheiten (meist Wörter) steht, nach denen gesucht wird, und n für die maximale Anzahl an Wörtern, die zwischen den beiden stehen dürfen (wobei es egal ist, ob String2 *vor* oder *nach* String1 auftritt - sucht man hingegen nach einer festen Wortreihenfolge, sollte man stattdessen Wortabstandsoperatoren wie #3 benutzen, vgl. die Hilfe zur Suche). Um Belege für *Sinn machen/ergeben* zu finden, kann man sich z.B. für einen Wortabstand von 3 entscheiden und die entsprechenden Treffer mit folgender Suchanfrage finden:

near($l=/ergeben|machen/g, $l=Sinn, 3)

Diese Suchanfrage ist wie folgt zu lesen: Finde alle Wörter, die auf der Lemma-Ebene (dafür steht $l) als *ergeben* oder *machen* lemmatisiert sind und denen im Abstand von 3 Wörtern vorher oder nachher ein als *Sinn* lemmatisiertes Wort folgt.

Fig. 1: Suchabfrage zu „Sinn machen“ vs. „Sinn ergeben“ in DWDS.

Wie bereits der erste Beleg in Fig. 1 zeigt, findet man so natürlich auch einige Fehltreffer – und je höher man den Wortabstand setzt, desto mehr Fehltreffer findet man natürlich auch. Hier gilt es, abzuwägen zwischen a) möglichst wenigen Fehltreffern auf die Gefahr hin, viele Treffer zu übersehen einerseits und b) einer möglichst erschöpfenden Suche, die u.U. eine Vielzahl von Fehltreffern mit sich bringt, deren Beseitigung für Sie erschöpfend sein kann. Wenn irgend möglich, wünschen wir uns natürlich eine möglichst exhaustive Suche – solange der Aufwand bei der manuellen Bereinigung der Daten sich in einem einigermaßen vertretbaren Rahmen hält, sollten Sie sich also für b) entscheiden.

3. Eine Konkordanz exportieren

Bevor die manuelle Bereinigung beginnen kann, müssen Sie die Daten zunächst exportieren. Genauer gesagt, exportieren Sie eine **Konkordanz**, also eine Belegliste, meist im Format „Key Word in Context“ (KWIC), dem wir im Folgenden immer wieder begegnen werden.

Exkurs: Die Konkordanz - Fluch oder Segen?

Dass es innerhalb der Korpuslinguistik sehr unterschiedliche methodische Ansätze geben kann, zeigt sich sehr schön in der Gegenüberstellung zweier Schlagwörter, die ich in zwei Vorträgen auf Konferenzen im Sommer 2017 aufgeschnappt habe: Während ein renommierter Korpuslinguist die Konkordanz als „weapon of first and last resort“ bezeichnete, plädierte ein anderer für „concordance-free corpus linguistics“.
Beide Auffassungen haben ein Stückweit ihre Berechtigung: Korpuslinguistik ohne Konkordanzen ist möglich und kann sinnvoll sein, beispielsweise indem man statt mit Belegen im Kontext einfach mit dem gesamten Korpus arbeitet – sei es, indem man es selber komplett liest (bei einigen Fragestellungen ist das unumgänglich; natürlich ist das nur bei sehr kleinen Korpora möglich) oder indem man es von einem Algorithmus komplett „lesen“ lässt. Letzteres ist z.B. beim semantic vector-space modelling der Fall – auch wenn dieses mit dem unmittelbaren Kontext der untersuchten Wörter und damit letztlich auch mit Konkordanzen arbeitet. Und wenn man ein Korpus komplett liest, dann in aller Regel auch, um Belege für ein bestimmtes Phänomen zu finden, das man sucht – und am Ende landet man dann auch wieder bei einer Belegliste, also einer Konkordanz. Letzten Endes ist Korpuslinguistik also doch auch immer „Konkordanzlinguistik“.
Kritik an einer „Konkordanzlinguistik“ ist allerdings dort berechtigt, wo Korpusbelege ohne systematische Analyse lediglich zu illustrativen Zwecken herangezogen werden oder wo Korpusbelege einer sich nicht auf handfeste Kriterien stützenden, intersubjektiv nicht nachprüfbaren Analyse unterzogen werden.

Das ist über den Button „Text exportieren“ möglich, den Sie in Fig. 1 etwas rechts unterhalb von der Mitte des Screenshots sehen. Es öffnet sich der in Fig. 2 dargestellte Exportdialog.

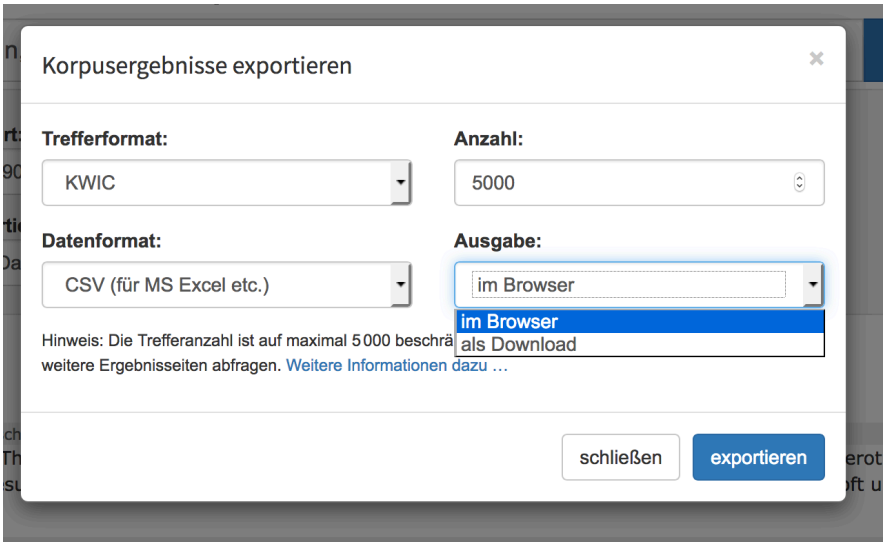

Fig. 2: Exportdialog im DWDS.

4. Die Konkordanz bereinigen, annotieren und analysieren
Um die Konkordanz zu bereinigen und sie ggf. mit weiteren Annotationen zu versehen, kann man die exportierte Datei nun mit der in Tutorial 1 („Grundlegendes“) im Abschnitt zu Excel und Calc dargelegten Methode in einem Tabellenkalkulationsprogramm öffnen. Im Folgenden illustriere ich es mit Hilfe des kostenlosen Calc. Im Importdialog erkennt Calc (zumindest auf meinem Rechner, s. Fig. 3) ganz richtig, dass es sich um eine Komma-separierte Datei mit Anführungszeichen als Textqualifizierer (*text delimiter*) handelt. Bei mir funktionierte lediglich die erste Spalte (Date) mit den Default-Einstellungen nicht, deshalb habe ich allen Spalten den Spaltentyp „Text“ zugewiesen (s. Fig. 3). Alternativ kann man der Datumsspalte natürlich auch gleich den Spaltentyp „Date“ zuweisen; wenn man die Daten später mit R weiterverarbeiten möchte, ist man aber ggf. mit „Text“ auf der sichereren Seite.

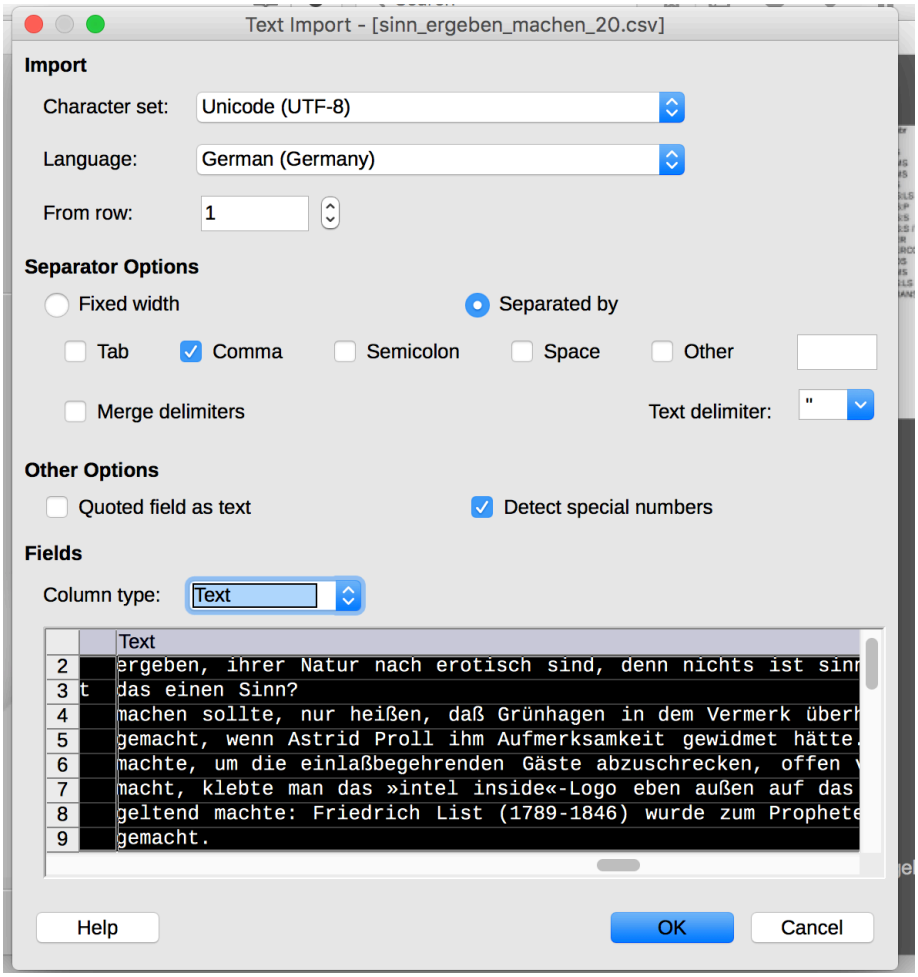

Fig. 3: Importdialog in Calc.

Um die weitere Bearbeitung zu vereinfachen, ist es oft sinnvoll, die Spalten etwas weniger breit zu machen. Dazu zoome ich mit dem Schieberegler unten rechts aus dem Dokument heraus und passe die Spaltengröße so an, dass ich die Spalten „ContextBefore“, „Hit“ und „ContextAfter“ auch dann noch zusammen auf dem Bildschirm sehe, wenn ich wieder so weit in das Dokument hineinzoome, dass der Text gut lesbar ist (Fig. 4).

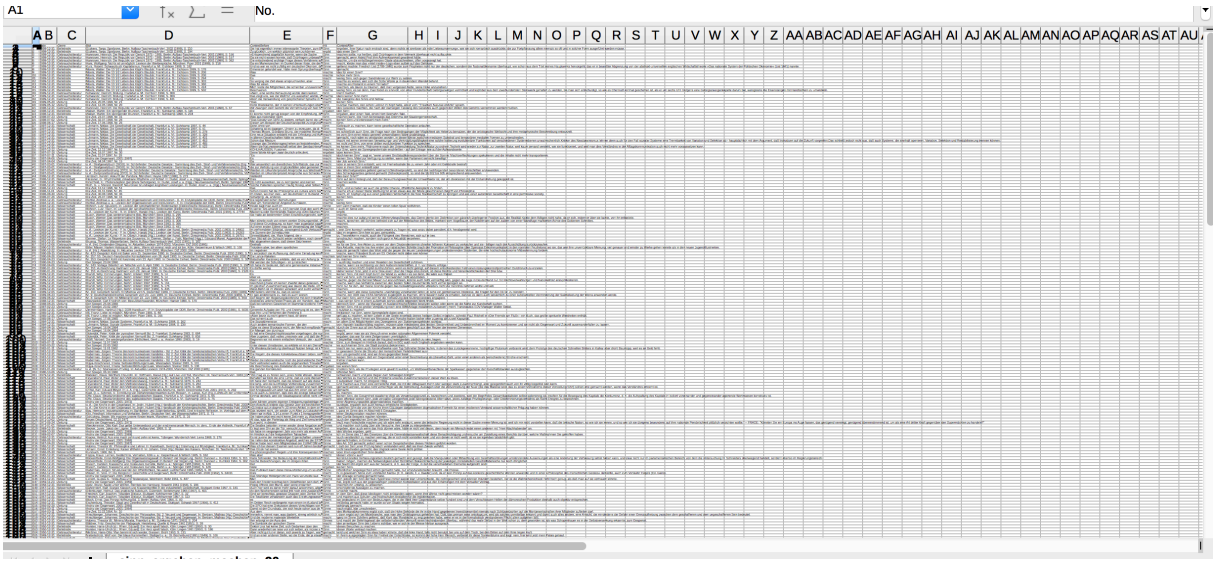

Fig. 4: Anpassen der Spaltenbreite in Calc.

Weiterhin kann es sehr hilfreich sein, die Spalte „ContextBefore“ rechtsbündig zu formatieren, sodass man beim Lesen der Konkordanz bequem vom linken Kontext über den Treffer zum rechten Kontext kommt. Wenn man ganz oben auf die Spalte klickt (dort, wo die Buchstaben stehen), wird die gesamte Spalte markiert. Darüber hinaus empfiehlt es sich in vielen Fällen, die Spalten, die für die manuelle Bearbeitung der Konkordanz nicht benötigt werden, auszublenden (**nicht** löschen - die Informationen können später evtl. noch wichtig werden!). Das geht, indem man die Spalte(n) markiert und dann Rechtsklick > Ausblenden wählt.

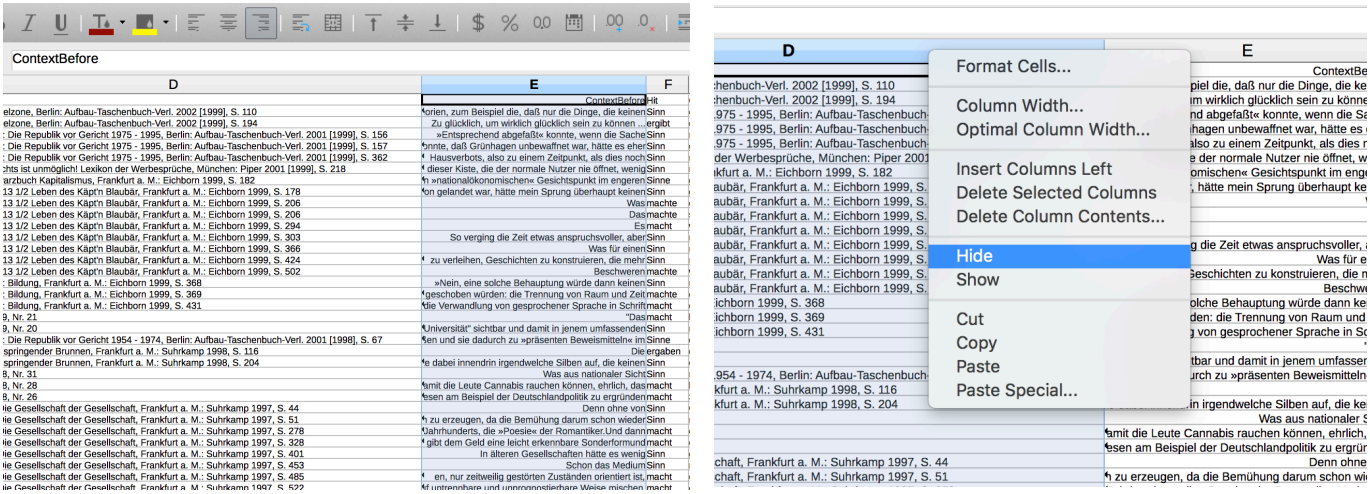

Fig. 5: Die Spalte „ContextBefore“ wird rechtsbündig formatiert; Spalten, die bei der manuellen Annotation nicht benötigt werden, werden ausgeblendet.