Einfache Korpusanalysen: Ein Schnelleinstieg

Stefan Hartmann

2019-06-09

Contents

1 Einstieg

Ziel dieses Tutorials ist es, Anfänger*innen einen möglichst niedrigschwelligen Einstieg in einfache Korpusanalysen zu ermöglichen. Es ist insbesondere für Studierende gedacht, die z.B. für eine Seminararbeit eine Korpusrecherche durchführen möchten, aber bislang noch keine praktische Erfahrung mit korpuslinguistischen Methoden sammeln konnten. Das Tutorial bietet anhand eines konkreten Beispiels eine Schritt-für-Schritt-Anleitung, wie man von der Fragestellung zur Datengewinnung hin zur Analyse der Daten gelangen kann.

Um wirklich einen Schnelleinstieg bieten zu können, muss ich notwendigerweise vieles vereinfachen. Für Ihre konkrete Korpusstudie werden Sie daher wahrscheinlich nicht umhinkommen, sich an der einen oder anderen Stelle tiefer einzulesen. Dafür verweise ich im Text immer wieder auf weiterführende Ressourcen. Teilweise finden sich auch in diesem Tutorial vertiefende Passagen, die Sie aufklappen können:

klick mich

Hallo, ich bin eine vertiefende Passage.

Sonst gibt es hier nichts zu sehen. Sie können mich gern wieder schließen. Danke.

2 Von der Fragestellung zur Konkordanz

Die meisten empirischen Studien lassen sich auf folgende Schritte herunterbrechen:

- Eine Fragestellung formulieren
- Daten erheben
- Daten auswerten.

2.1 Eine Fragestellung formulieren

Der erste Schritt ist wahrscheinlich der wichtigste. Nur wenn Sie eine gute Forschungsfrage haben, können Sie eine aussagekräftige empirische Analyse durchführen. Aus der Forschungsfrage ergibt sich die Methode: Für manche Fragestellungen bietet sich z.B. eine Fragebogenstudie an, für eine eine psycho- oder neurolinguistische Herangehensweise, für wieder andere eine Korpusrecherche.

Das heißt auch: Wenn Sie eine Korpusanalyse durchführen möchten, brauchen Sie eine Fragestellung, die korpuslinguistisch operationalisierbar ist. Beispielsweise lässt sich eine Frage wie "Welche Gehirnareale werden beim Hören von Bewegungsverben aktiviert?" natürlich nicht mit Hilfe von Korpusdaten beantworten.

Für unsere Beispielanalyse werfen wir einen Blick auf die prädikative Verwendung der Partizipien *programmiert* und *vorprogrammiert*. Letzteres ist manchen Sprachpflegern ein Dorn im Auge: So bezeichnet es Batian Sick als

> "umgangssprachliches Blähwort, über das schon Heerscharen von Sprachpflegern hergefallen sind – vergebens, denn es wird immer munter weiter vorprogrammiert. Dabei wissen nicht nur Programmierer: Man programmiert immer im Voraus, die Vorsilbe vor- ist daher pleonastisch, zu Deutsch: doppelt gemoppelt." — https://bastiansick.de/kolumnen/abc/vorprogrammiertprogrammiert/

Was Sprachpfleger wie Sick jedoch oft verkennen, ist, dass Sprache nicht immer "logisch" ist. Vielmehr suchen sich Wörter oft eigene Nischen. Beispielsweise ist mein Bürostuhl kein *Rollstuhl*, obwohl er Rollen hat - denn das Wort *Rollstuhl* hat eine eigene Bedeutung angenommen, die sich nicht kompositional aus seinen Einzelteilen ergibt. Im Falle von *vorprogrammiert* hingegen passt zwar die Paraphrase 'im Voraus programmiert'. Aber trotzdem wäre denkbar, dass das Wort eine Spezialisierung erfahren hat: Wird *programmiert* möglicherweise eher dann verwendet, wenn ein Programmierungsvorgang im wörtlichen Sinn gemeint ist, und *vorprogrammiert* eher dann, wenn ein z.B. ein Skandal oder eine Katastrophe "vorprogrammiert" sind? Das ist die Fragestellung, der wir im Folgenden nachgehen möchten.

Fragestellungen und Hypothesen

Die Unterscheidung von **Fragestellung** und **Hypothese** bereitet Anfänger*innen oft Schwierigkeiten. Beide hängen eng zusammen. In unserem Beispiel könnte man die Frage in eine Hypothese umformulieren: "vorprogrammiert wird eher in metaphorischem und programmiert eher im wörtlichen Sinn verwendet."

Hypothesen ergeben sich in der Regel aus konkreten Fragestellungen. Beispielsweise könnte in einer soziologischen oder politikwissenschaftlichen Studie die Fragestellung lauten: Welchen Einfluss hat das Alter auf das Wahlverhalten in Deutschland? Da man zu diesem Themengebiet aus der bisherigen Forschung und aus der Alltagserfahrung das eine oder andere schon weiß, kann man begründete Annahmen darüber treffen, wie die Antwort auf diese Frage aussieht. So könnte man davon ausgehen, dass z.B. ältere Menschen eher etablierte und vielleicht auch eher konservative Parteien wählen und dass außerdem bei Älteren eine höhere Wahlbeteiligung vorliegt. Diese Annahmen nennt man Hypothesen. Sie werden auf Grundlage der Daten, die man erhebt, überprüft.

Nicht immer ist es möglich oder notwendig, konkrete Hypothesen zu formulieren. Gerade bei Phänomenen, über die noch sehr wenig bekannt ist, bietet es sich manchmal an, **explorativ**, also "erkundend", zu arbeiten. Auch dann gehe ich mit einer Fragestellung an meine Daten heran, ohne jedoch im Voraus eine Erwartung zu haben, wie die Antwort auf meine Frage aussehen wird.

2.2 Daten erheben

2.2.1 Suchsyntax

Für die Datenerhebung verwenden wir das DWDS-Kernkorpus des 20. Jahrhunderts, das über dwds.de zugänglich ist. Wir suchen auf der Wortebene mit Hilfe von regulären Ausdrücken nach den Formen *programmiert* und *vorprogrammiert*. Dafür benutzen wir den Suchstring `@programmiert || @vorprogrammiert`. Das @-Zeichen bedeutet, dass wir genau diese Strings suchen und keine anderen Wortformen wie *programmierte*, *programmiertes* etc. Da uns nur die prädikative Verwendung interessiert, brauchen wir die flektierten Wortformen nicht. Der horizontale Strich | ist der ODER-Operator; dass man ihn hier doppelt setzen muss, ist eine Besonderheit der DWDS-Suchsyntax.

Alternative Suchabfrage mit regulären Ausdrücken Alternativ können wir das gleiche Ergebnis auch durch Verwendug regulärer Ausdrücke erzielen: `$w=/(vor)?programmiert/g`. Ich ermutige alle, die sich mit

Korpuslinguistik beschäftigen wollen, sehr, sich mit regulären Ausdrücken vertraut zu machen. Allerdings unterstützt die DWDS-Suchsyntax reguläre Ausdrücke derzeit nur in sehr beschränktem Maße. (Deutlich besser ist in dieser Hinsicht das alternative Abfrageportal Dstar, das jedoch für Anfänger*innen nur bedingt geeignet ist.)

Zur Suche im DWDS und anderswo - Die Hilfe zur Suche im DWDS findet sich hier.

- Einen Einstieg in reguläre Ausdrücke bietet z.B. regular-expressions.info.
- In den Begleitmaterialien zu meiner "Deutschen Sprachgeschichte" finden sich ebenfalls einige Tutorials zur Suche in einschlägigen Korpora.
- Sehr empfehlenswert und erfreulich ausführlich ist außerdem die Korpuslinguistik-Seite von Noah Bubenhofer.

2.2.2 Export

Die Suche liefert uns 88 Treffer, die nun im Browser in ihrem jeweiligen Kontext dargestellt werden. Diese Daten wollen wir nun exportieren, und zwar im "Key Word in Context" (KWIC)-Format. Damit ist gemeint, dass der Suchtreffer zusammen mit seinem unmittelbaren Kontext dargestellt wird. Erfreulicherweise bietet das DWDS eine sehr gute Exportfunktion, die es erlaubt, Daten im CSV-Format zu speichern.

Eine solche Sammlung von Korpusbelegen, wie wir sie jetzt exportiert haben, nennt man in der Korpuslinguistik **Konkordanz**. Der Formatname "CSV" steht für "Comma-Separated Values". Das heißt, in der Datei sind die einzelnen Werte durch Kommata voneinander abgetrennt. In einem Texteditor sieht das Ganze so aus wie in 2. Wie Sie sehen, enthält die Datei neben den Korpusbelegen selbst auch Metadaten zu den einzelnen Belegen, z.B. zu Autor*in, Titel etc.

Damit können wir zunächst noch wenig anfangen: Wir wollen die Konkordanz in ein Tabellenkalkulationsprogramm einlesen.

2.2.3 Import in ein Tabellenkalkulationsprogramm

Wenn Sie Microsoft Excel auf Ihrem Rechner installiert haben, sind die Default-Einstellungen höchstwahrscheinlich so gesetzt, dass CSV-Dateien in Excel geöffnet werden, wenn Sie darauf doppelklicken. Warum das keine gute Idee ist, zeigt der folgende Screenshot 3 (rote Hervorhebungen von mir nachträglich hinzugefügt).

Hier sind einige Sonderzeichen verlorengegangen, weil Excel die Kodierung der Datei nicht richtig erkannt hat. Es gibt mehrere Wege, diesem Problem zu begegnen. Ich empfehle hier zwei: Einen für Excel und einen für die freie Alternative Calc.

2.2.3.1 Import in Excel

1. Öffnen Sie die Datei in einem Texteditor. Für Windows empfehle ich Notepad++, für Mac die kostenlose (und für unsere Zwecke völlig ausreichende) Version von BBEdit, für Linux gibt es z.B. Notepadqq.
2. Markieren Sie mit Strg+A bzw. Cmd+A den gesamten Text.
3. Öffnen Sie ein leeres Tabellenblatt in Excel. Die nächsten Schritte, 4 bis , sind in 4 visualisiert.
4. In den meisten Fällen sollten Sie nun einfach mit Strg+V bzw. Cmd+V die Daten einfügen könnn. In manchen Fällen müssen Sie jedoch, wie im Screencast 4, die Option "Paste Special" verwenden (dt. "Inhalte einfügen") und angeben, dass Sie den Unicode-Text einfügen möchten.
5. Mit Klick auf das kleine Klemmbrett-Symbol gelangen Sie zum Textimport-Assistenten. Hier müssen Sie Excel sagen, wie der eingefügte Text strukturiert ist. Auf der ersten Seite sagen Sie, dass es sich um einen Text handelt, bei dem die einzelnen Spalten durch ein Trennzeichen getrennt sind ("Delimited") -

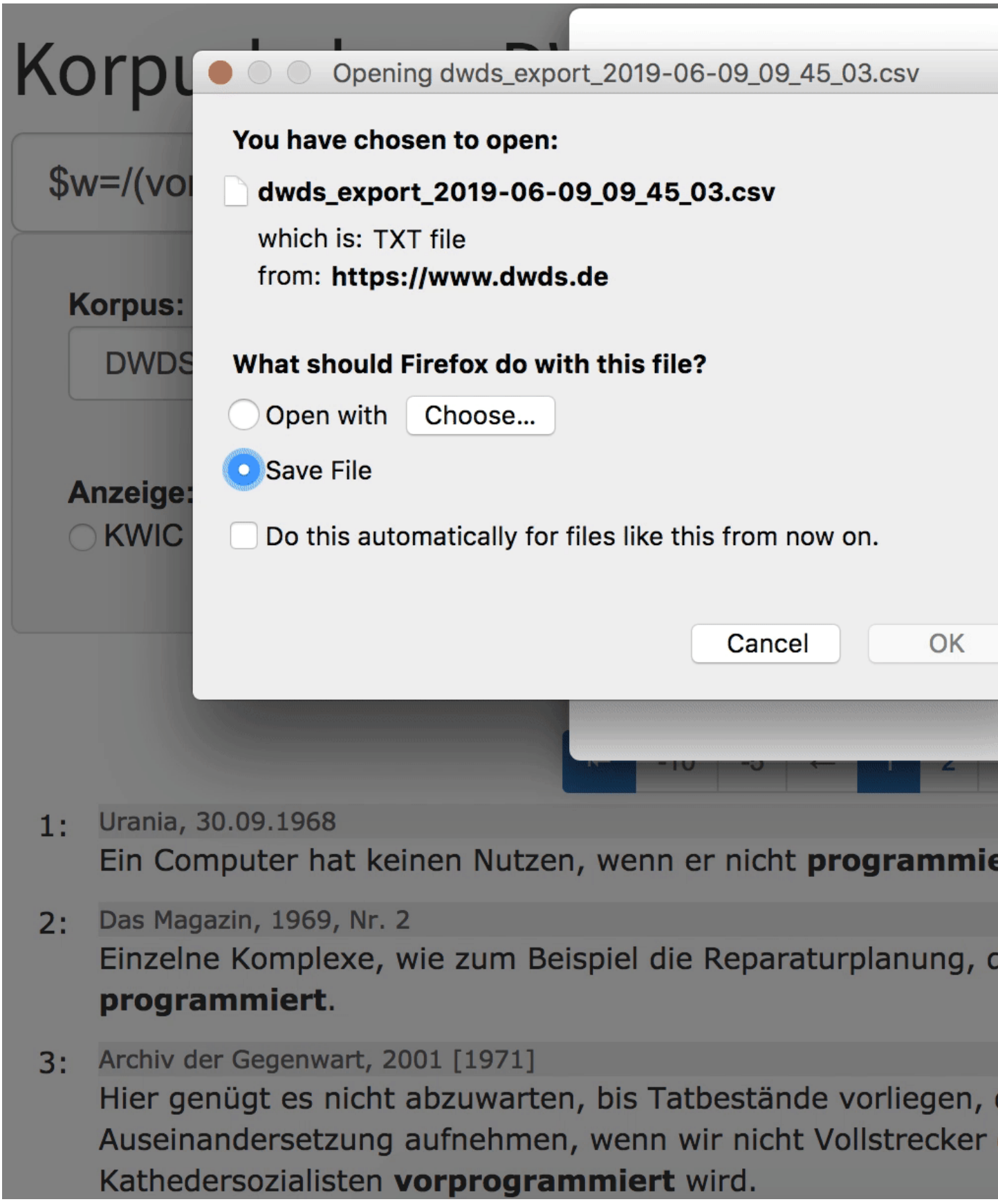
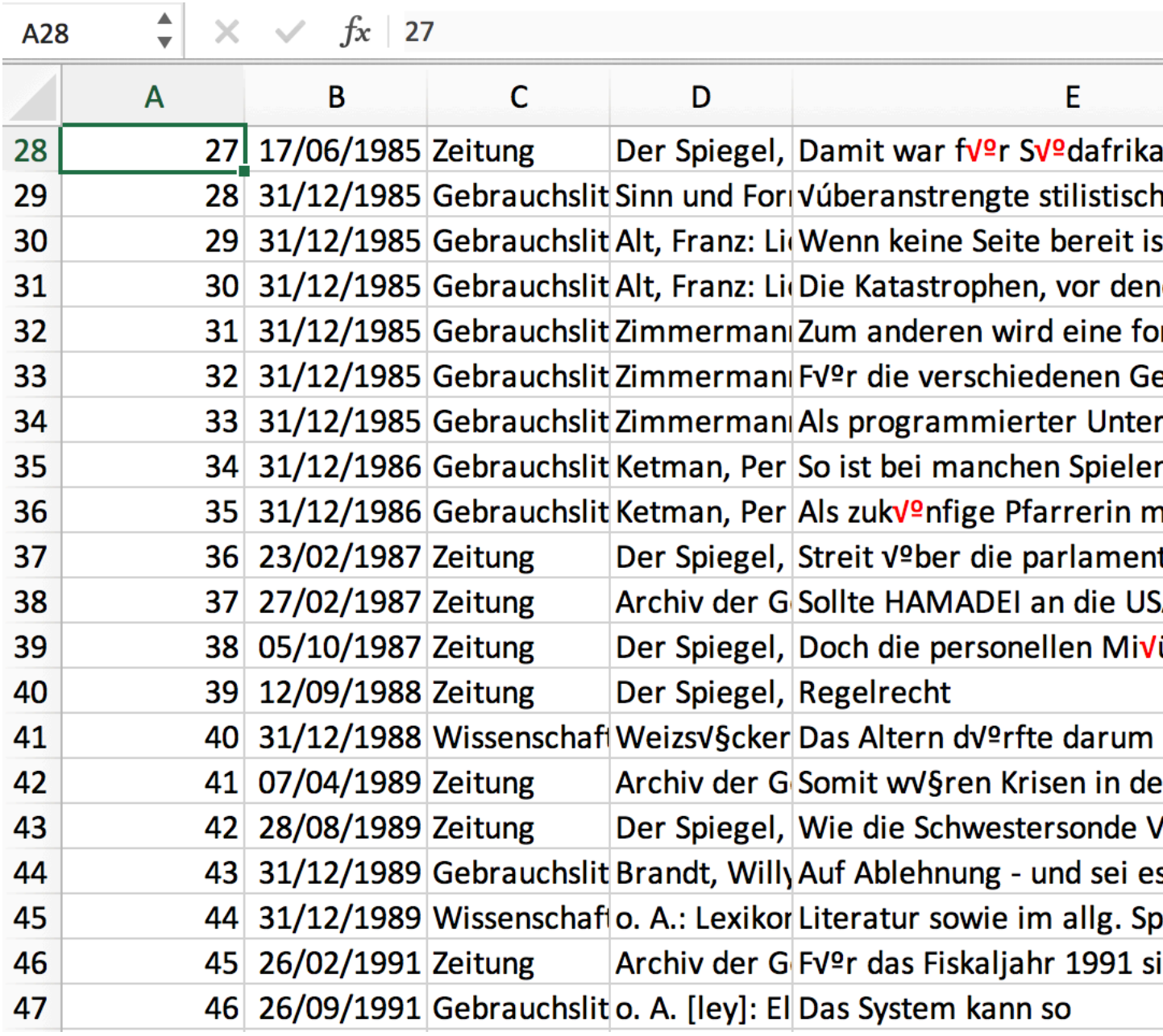


Figure 1: Export aus dem DWDS

Figure 2: Konkordanz im Texteditor



	A	B	C	D	E
28	27	17/06/1985	Zeitung	Der Spiegel,	Damit war fV$r SV$dafrika
29	28	31/12/1985	Gebrauchslit	Sinn und For	Vúberanstrengte stilistisch
30	29	31/12/1985	Gebrauchslit	Alt, Franz: Li	Wenn keine Seite bereit is
31	30	31/12/1985	Gebrauchslit	Alt, Franz: Li	Die Katastrophen, vor den
32	31	31/12/1985	Gebrauchslit	Zimmerman	Zum anderen wird eine fo
33	32	31/12/1985	Gebrauchslit	Zimmerman	FV$r die verschiedenen Ge
34	33	31/12/1985	Gebrauchslit	Zimmerman	Als programmierter Unter
35	34	31/12/1986	Gebrauchslit	Ketman, Per	So ist bei manchen Spieler
36	35	31/12/1986	Gebrauchslit	Ketman, Per	Als zukV$nfige Pfarrerin m
37	36	23/02/1987	Zeitung	Der Spiegel,	Streit V$ber die parlament
38	37	27/02/1987	Zeitung	Archiv der G	Sollte HAMADEI an die US
39	38	05/10/1987	Zeitung	Der Spiegel,	Doch die personellen MiV
40	39	12/09/1988	Zeitung	Der Spiegel,	Regelrecht
41	40	31/12/1988	Wissenschaf	WeizsV§cker	Das Altern dV$rfte darum
42	41	07/04/1989	Zeitung	Archiv der G	Somit wV§ren Krisen in de
43	42	28/08/1989	Zeitung	Der Spiegel,	Wie die Schwestersonde V
44	43	31/12/1989	Gebrauchslit	Brandt, Willy	Auf Ablehnung - und sei es
45	44	31/12/1989	Wissenschaf	o. A.: Lexikor	Literatur sowie im allg. Sp
46	45	26/02/1991	Zeitung	Archiv der G	FV$r das Fiskaljahr 1991 si
47	46	26/09/1991	Gebrauchslit	o. A. [ley]: El	Das System kann so

Figure 3: Konkordanz bei direktem Öffnen in Excel

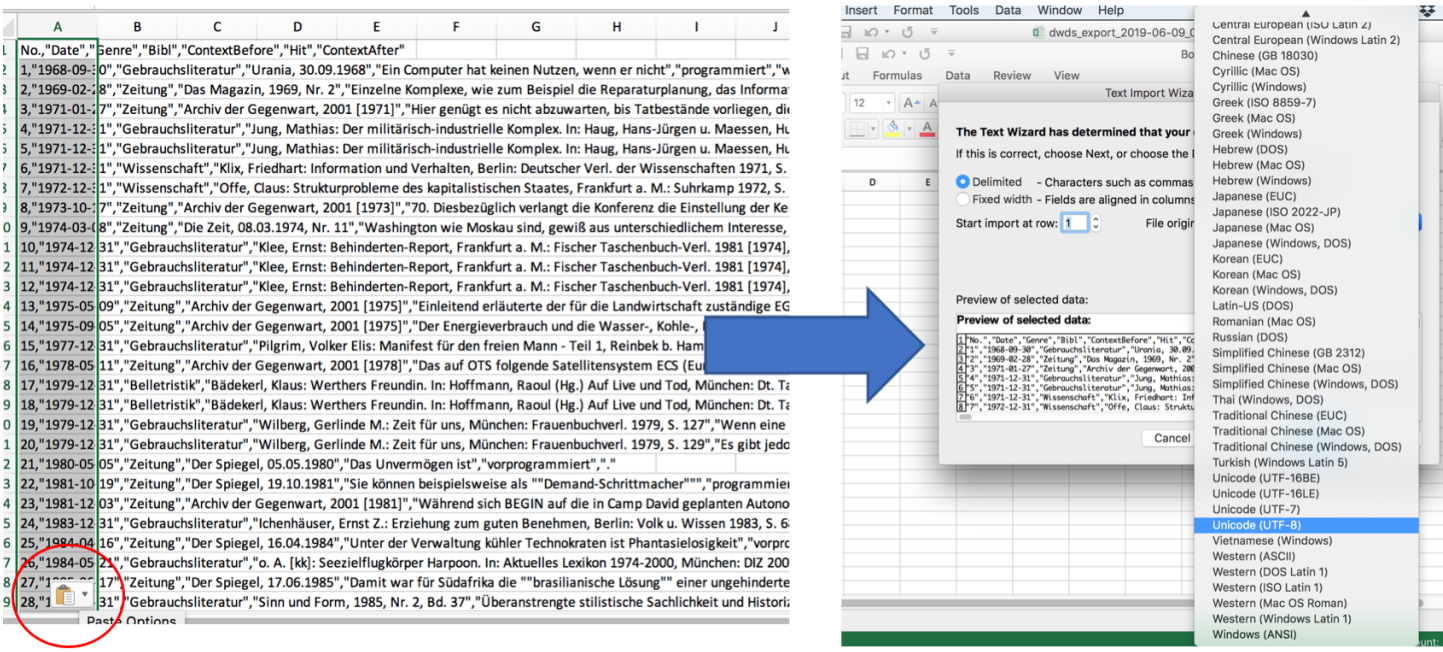


Figure 4: Import in Excel

diese Option ist in der Regel schon angewählt. Außerdem teilen Sie Excel hier mit, dass der eingefügte Text UTF-8-formatiert ist.

6. Auf der nächste Seite des Textimport-Assistenten geben Sie an, dass Kommata als Spaltentrenner benutzt werden. Bei den Textqualifizierern müssen Sie nichts ändern, da hier schon Anführungszeichen ausgewählt sind: Wie Sie in 2 sehen können, werden Anführungszeichen in der CSV-Datei genutzt, um zusammengehörigen Text zusammenzuhalten (denn wären sie nicht da, würde Excel jedes Komma im Text für einen Spaltentrenner handeln.)

7. Dieser letzte Schritt erübrigt sich meistens, kann aber nicht schaden: Zuletzt können Sie noch alle Spalten als “Text” formatieren. (Die Datumsspalte können Sie prinzipiell auch als “Datum” formatieren, falls Sie ausschließlich in Excel weiterarbeiten, aber tendenziell rate ich davon ab - gerade bei einer späteren Konversion in andere Dateiformate kann dabei alles mögliche schiefgehen…)

2.2.3.2 Import in Calc

Öffnet man die Datei im kostenlosen Tabellenkalkulationsprogramm Calc von LibreOffice (mit Rechtsklick > Öffen mit), so öffnet sich zunächst automatisch der Textimportassistent. Hier muss man Calc mitteilen, welches Format die Datei hat. In unserem Fall ist der Text UTF-8-kodiert, wir haben Kommas als Spaltentrenner und Anführungszeichen als Textqualifizieren, wie in 5.

3 Von der Konkordanz zur Analyse

Nun haben wir die Konkordanz erfolgreich in ein Tabellenkalkulationsprogramm importiert. Hier können wir beliebig viele weitere Spalten hinzufügen. Das können wir nutzen, um die exportierten Belege mit **Annotationen** zu versehen.

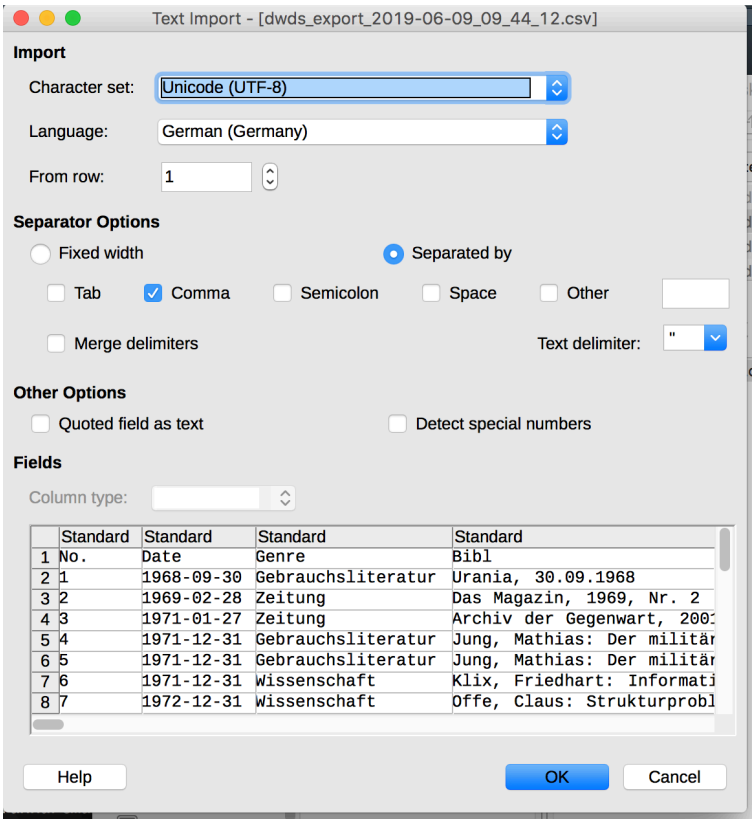


Figure 5: Import in Calc

3.1 Annotation

Versieht man Daten mit zusätzlichen Informationen, so nennt man diesen Prozess Annotation. In der Korpuslinguistik stellt die Annotation einen ganz wesentlichen Schritt dar, der gewissermaßen die Brücke schlägt von der qualitativ-philologischen Analyse einzelner Belege zur quantitativen Auswertung.

Wir nutzen im Folgenden die Annotation, um unsere Daten in Kategorien zu unterteilen, die für unsere Fragestellung sinnvoll sind. Dafür müssen wir uns zunächst darüber im Klaren sein, was wir von unseren Daten überhaupt wissen wollen, d.h. wir müssen unsere eingangs genannte Fragestellung operationalisieren.

Zur Erinnerung: Unsere Fragestellung lautet, ob bei prädikativem Gebrauch *vorprogrammiert* gegenüber *programmiert* bevorzugt wird, wenn es sich um einen metaphorischen Kontext handelt.

Konkret bedeutet das, dass wir für jeden Datenpunkt folgende Fragen beantworten müssen:

1. Handelt es sich um eine prädikative Verwendung? - Schon ein kurzer Blick auf die Daten zeigt, dass sich notwendigerweise einige **Fehltreffer** eingeschlichen haben: Häufig finden sich z.B. Passivkonstruktionen wie *Es gibt jedoch medizinische Gründe, aus denen eine Geburt eingeleitet oder sogar programmiert werden muß.* Uns interessieren aber nur Fälle, in denen das Partizip selbst das Prädikat bildet, also z.B. *Der Computer ist programmiert* und *Die Katastrophe war vorprogrammiert.*
2. Handelt es sich um eine metaphorische Verwendung?