

Einfache Korpusanalysen: Ein Schnelleinstieg

Stefan Hartmann

2019-06-14

Inhaltsverzeichnis

1 Einstieg	1
2 Von der Fragestellung zur Konkordanz	2
2.1 Eine Fragestellung formulieren	2
2.2 Daten erheben	3
2.2.1 Suchsyntax	3
2.2.2 Export	4
2.2.3 Import in ein Tabellenkalkulationsprogramm	4
2.2.3.1 Import in Excel	5
2.2.3.2 Import in Calc	6
3 Von der Konkordanz zur Analyse	6
3.1 Annotation	6
3.1.1 Annotation prädikativ vs. nicht-prädikativ	7
3.1.1.1 Umsetzung in Excel	8
3.1.1.2 Umsetzung in Calc	10
3.1.2 Annotation metaphorisch vs. nicht-metaphorisch	11
3.2 Auswertung und Visualisierung	12
3.2.1 Auswertung und Visualisierung in Excel	13
3.2.2 Auswertung und Visualisierung in Calc	17
4 Und nun...?	19
4.1 Interpretation und Einordnung	19
4.2 Methodenkritik und offene Fragen	20

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. Dabei nutzen wir folgende Ressourcen bzw. Programme:

- Das Kernkorpus des 20. Jahrhunderts des Digitalen Wörterbuchs der Deutschen Sprache, verfügbar über dwds.de.
- ein Tabellenkalkulationsprogramm, wobei alle wesentlichen Arbeitsschritte sowohl für Microsoft Excel als auch für die kostenlose Alternative LibreOffice Calc beschrieben werden. Es genügt natürlich, wenn Sie mit einem der beiden Programme arbeiten; die Abschnitte zum jeweils anderen Programm können Sie dann getrost überspringen.

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 gelegentlich auf weiterführende Ressourcen. Teilweise finden sich auch in diesem Tutorial vertiefende Passagen, die Sie (in der HTML-Version) aufklappen können:

klick mich

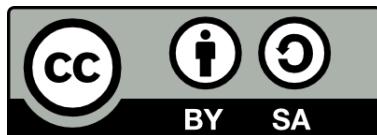
Hallo, ich bin eine vertiefende Passage.

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

Ein Hinweis vorab: Das Tutorial setzt keine Kenntnisse in der Korpuslinguistik oder im Umgang mit Tabellenkalkulationsprogrammen voraus, wohl aber grammatische Grundkenntnisse. Sollten Sie die Fachbegriffe nicht verstehen, empfehle ich sehr, sie nachzuschlagen und die entsprechenden Wissenslücken zu schließen.

Und noch ein Hinweis: Das Tutorial liegt sowohl im HTML-Format als auch im PDF-Format vor (die PDF-Version können Sie im HTML-Dokument durch Klick auf den PDF-Button ganz oben herunterladen). Die HTML-Version arbeitet viel mit animierten GIFs, die in der PDF-Version natürlich nicht zu sehen sind. Dafür benötigt die PDF-Version weit weniger Speicherplatz.

Ein letzter Hinweis: Da ich aus Gründen, die hier auszuführen zu viel Platz wegnehmen würde, mit den englischen Versionen von Excel und Calc arbeite, sehen Sie in den Screenshots und Screencasts immer die englischen Varianten der einschlägigen Befehle. Im Text nenne ich teilweise die deutsche Übersetzung, ohne aber in jedem Einzelfall geprüft zu haben, ob das auch wirklich die Übersetzung ist, die in den deutschen Versionen der Programme verwendet wird.



Dieses Tutorial wurde mit Hilfe von Bookdown für R geschrieben und publiziert. Es ist lizenziert unter CC-BY-SA und kann gerne mit Quellenangabe weitergegeben und adaptiert werden.

Zitiervorschlag: Hartmann, Stefan. 2019. Einfache Korpusanalysen: Ein Schnelleinstieg. <https://empirical-linguistics.github.io/korpus-schnelleinstieg/>. doi 10.5281/zenodo.3246336.

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 andere 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 prädiktative 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 erkennen, 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 kompositionsl 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 prädiktative 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 Verwendung 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

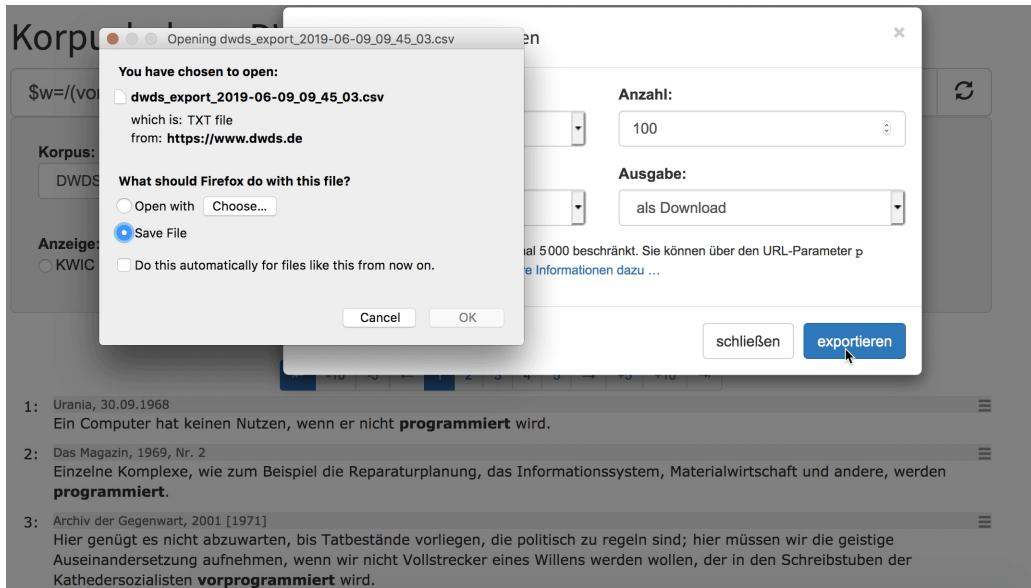


Abbildung 1: Export aus dem DWDS

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 Bubenhofen.

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.

Users > stefanhartmann > Dropbox > Privat > Tutorials > korpus-schnelleinstieg > data > dwds_export_2019-06-09_09_44_12.csv

```

1 "No.,""Date","Genre","Bibl","ContextBefore","Hit","ContextAfter"
2 "1","1968-09-30","Gebrauchsleiteratur","Urania 30.09.1968","Ein Computer hat keinen Nutzen, wenn er nicht","programmiert","wird."
3 "2","1969-02-28","Zeitung","Das Magazin, 1969, Nr. 2","Einzelne Komplexe, wie zum Beispiel die Reparaturplanung, das Informationssystem, Mater.
4 "3","1971-01-27","Zeitung","Archiv der Gegenwart, 2001 [1971]","Hier genügt es nicht abzuwarten, bis Tatbestände vorliegen, die politisch zu ri
5 "4","1971-12-31","Gebrauchsleiteratur","Jung, Mathias: Der militärisch-industrielle Komplex. In: Haug, Hans-Jürgen u. Maessen, Hubert (Hgg.) Kr.
6 "5","1971-12-31","Gebrauchsleiteratur","Jung, Mathias: Der militärisch-industrielle Komplex. In: Haug, Hans-Jürgen u. Maessen, Hubert (Hgg.) Kr.
7 "6","1971-12-31","Gebrauchsleiteratur","Klix, Friedhart: Information und Verhalten, Berlin: Deutscher Verl. der Wissenschaften 1971, S. 731","Diese M
8 "7","1972-12-31","Wissenschaft","Offe, Claus: Strukturprobleme des kapitalistischen Staates, Frankfurt a. M.: Suhrkamp 1972, S. 90","Das bedeutet
9 "8","1973-10-17","Zeitung","Archiv der Gegenwart, 2001 [1973]","70. Diesbezüglich verlangt die Konferenz die Einstellung der Kernversuche, die
10 "9","1974-03-08","Zeitung","Die Zeit, 08.03.1974, Nr. 11","Washington und Moskau sind, gewiß aus unterschiedlichen Interesse, auf einen israel.
11 "10","1974-12-31","Gebrauchsleiteratur","Klee, Ernst: Behinderten-Report, Frankfurt a. M.: Fischer Taschenbuch-Verl. 1981 [1974], S. 38","Die El
12 "11","1974-12-31","Gebrauchsleiteratur","Klee, Ernst: Behinderten-Report, Frankfurt a. M.: Fischer Taschenbuch-Verl. 1981 [1974], S. 93","Auch i
13 "12","1974-12-31","Gebrauchsleiteratur","Klee, Ernst: Behinderten-Report, Frankfurt a. M.: Fischer Taschenbuch-Verl. 1981 [1974], S. 94","Was ai
14 "13","1975-05-09","Zeitung","Archiv der Gegenwart, 2001 [1975]","Einleitend erläuterte der für die Landwirtschaft zuständige EG-Kommissar Petri
15 "14","1975-09-05","Zeitung","Archiv der Gegenwart, 2001 [1975]","Der Energieverbrauch und die Wasser-, Kohle-, Erdöl- und Gasreserven des Landes
16 "15","1977-12-31","Gebrauchsleiteratur","Pilgrim, Volker Elis: Manifest für den freien Mann - Teil 1, Reinbek b. Hamburg: Rowohlt 1983 [1977], '
17 "16","1978-05-11","Zeitung","Archiv der Gegenwart, 2001 [1978]","Das auf OTS folgende Satellitensystem ECS (European Communications Satellites
18 "17","1979-12-31","Gebrauchsleiteratur","Bädekerl Klaus: Werthers Freundin. In: Hoffmann, Raoul (Hg.) Auf Live und Tod, München: Dt. Taschenbuch-Ver
19 "18","1979-12-31","Gebrauchsleiteratur","Bädekerl Klaus: Werthers Freundin. In: Hoffmann, Raoul (Hg.) Auf Live und Tod, München: Dt. Taschenbuch-Ver
20 "19","1979-12-31","Gebrauchsleiteratur","Wilberg, Gerlinde M.: Zeit für uns, München: Frauenbuchverl. 1979, S. 127","Wenn eine Geburt auf sieben
21 "20","1979-12-31","Gebrauchsleiteratur","Wilberg, Gerlinde M.: Zeit für uns, München: Frauenbuchverl. 1979, S. 129","Es gibt jedoch medizinisch
22 "21","1980-05-05","Zeitung","Der Spiegel, 05.05.1980","Das Unvermögen ist","vorprogrammiert","."
23 "22","1981-10-19","Zeitung","Der Spiegel, 19.10.1981","Sie können beispielweise als ""Demand-Schrittmacher"" programmiert, werden, die ers
24 "23","1981-12-03","Zeitung","Archiv der Gegenwart, 2001 [1981]","Während sich BEGIN auf die in Camp David geplanten Autonomieverhandlungen ver
25 "24","1983-12-31","Gebrauchsleiteratur","Ichenhäuser, Ernst Z.: Erziehung zum guten Benehmen, Berlin: Volk u. Wissen 1983, S. 68","Ist es zu scl
26 "25","1984-04-16","Zeitung","Der Spiegel, 16.04.1984","Unter der Verwaltung kühler Technokraten ist Phantasielosigkeit", vorprogrammiert,"."
27 "26","1984-05-21","Gebrauchsleiteratur","o. A. (Khl): Seeschildkröter Harpoon. In: Aktuelles Lexikon 1974-2000, München: DIZ 2000 [1984]","Die
28 "27","1985-06-17","Zeitung","Der Spiegel, 17.06.1985","Damit war für Südafrika die ""brasiliatische Lösung"" einer ungehinderten Vermischung d
29 "28","1985-12-31","Gebrauchsleiteratur","Sinn und Forst, Franz: Überanstrengte stilistische Sachlichkeit und Historizität sowie ai
30 "29","1985-12-31","Gebrauchsleiteratur","Alt, Franz: Liebe ist möglich, München: Piper 1985, S. 129","Wenn keine Seite bereit ist, den ersten ko
31 "30","1985-12-31","Gebrauchsleiteratur","Alt, Franz: Liebe ist möglich, München: Piper 1985, S. 175","Die Katastrophen, vor denen wir heute ste
32 "31","1985-12-31","Gebrauchsleiteratur","Zimmermann, Hartmut (Hg.): DDR-Handbuch - F: In: Enzyklopädie der DDR, Berlin: Directmedia Publ. 2000
33 "32","1985-12-31","Gebrauchsleiteratur","Zimmermann, Hartmut (Hg.): DDR-Handbuch - M: In: Enzyklopädie der DDR, Berlin: Directmedia Publ. 2000
34 "33","1985-12-31","Gebrauchsleiteratur","Zimmermann, Hartmut (Hg.): DDR-Handbuch - U: In: Enzyklopädie der DDR, Berlin: Directmedia Publ. 2000

```

Abbildung 2: Konkordanz im Texteditor

A	B	C	D	E	F	G	H	I	J	K	L
28	27 17/06/1985	Zeitung	Der Spiegel, Damit war fv9r Sv9r Afrika die "brasiliare programmie.								
29	28 31/12/1985	Gebrauchsleiteratur	Gebrauchsleiteratur Sinn und Forvüberanstrenzte stilistische Sachlichkeit programmie ist.								
30	29 31/12/1985	Gebrauchsleiteratur	Gebrauchsleiteratur Alt, Franz: LiiWenn keine Seite bereit ist, den ersten vorprogrammier								
31	30 31/12/1985	Gebrauchsleiteratur	Gebrauchsleiteratur Alt, Franz: LiiDie Katastrophen, vor denen wir heute programmie #NAME?								
32	31 31/12/1985	Gebrauchsleiteratur	Zimmermann Zum anderen wird eine fortdauernde Div vorprogrammier.								
33	32 12/09/1985	Gebrauchsleiteratur	Zimmermann Für verschiedene Gebiete des ML programmie (vgl. fv9r für die gegenwv5rtig gv9!tige Planperiode:								
34	33 31/12/1985	Gebrauchsleiteratur	Zimmermann Als programmierter Unterricht (PU) wi programmie sind und der einem im Lehrprogramm gespeicherten Lehralgorithmus folgt, der die								
35	34 31/12/1986	Gebrauchsleiteratur	Gebr Ketman, Per So ist in manchen Spielen die Konfront vorprogrammier.								
36	35 31/12/1986	Gebrauchsleiteratur	Gebr Ketman, Per Als zuv9nfige Pfarrerin my9Vite Miria programmie, noch ehe es v9berhaupt mv9dig ist: ein Dilemma, das keine einfache Lv9sung ke								
37	36 23/02/1987	Zeitung	Der Spiegel, Streit v9ber die parlamentarische Salon programmie:								
38	37 27/02/1987	Zeitung	Archiv der G Sollte HAMADEI an die USA ausgeliefert vorprogrammie zudem wv9re dies das Todesurteil fv9r den Entf9!rten.								
39	38 05/10/1987	Zeitung	Der Spiegel, Doch die personellen Mv9griffe waren programmie.								
40	40 12/09/1988	Zeitung	Der Spiegel, Regelfrech programmie und in Szene gesetzt von staatlichen Instanzen war vor dem 9. November 1938 kein								
41	41 31/12/1988	Wissenschaft	Weizsäcker Das Altern dv9rfte darum genetisch programmie sein.								
42	42 07/04/1988	Zeitung	Archiv der G Somit wv9ren Krisen in den kommende vorprogrammier.								
43	43 28/08/1988	Zeitung	Der Spiegel, Wie die Schwestern Voyager 1 war programmie.								
44	44 31/12/1989	Gebrauchsleiteratur	Brandt, Willy Auf Ablehnung - und sei es nur, dav9 au programmie, die Bv9ndnisfreiheit fv9r Deutschland keinesfalls in Erwv9gung ziehen wollten.								
45	45 26/02/1991	Zeitung	Archiv der G Fv9r das Fiskaljahr 1991 sind Ausgaben programmie ist, das etwa dem Schuldendienst entspricht.								
46	46 26/09/1991	Gebrauchsleiteratur	o. [ley]: El Das System kann so programmie werden, dav9 es beispielweise Alarm gibt, wenn sich das Schiff einem Hindernis a								

Abbildung 3: Konkordanz bei direktem Öffnen in Excel

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

- 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.
- Markieren Sie mit Strg+A bzw. Cmd+A den gesamten Text.
- Öffnen Sie ein leeres Tabellenblatt in Excel. Die nächsten Schritte, 4 bis 7, sind in 4 visualisiert.
- In den meisten Fällen sollten Sie nun einfach mit Strg+V bzw. Cmd+V die Daten einfügen können. 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.
- Mit Klick auf das kleine Klemmbrett-Symbol gelangen Sie zum Textimport-Assistenten. Hier müssen

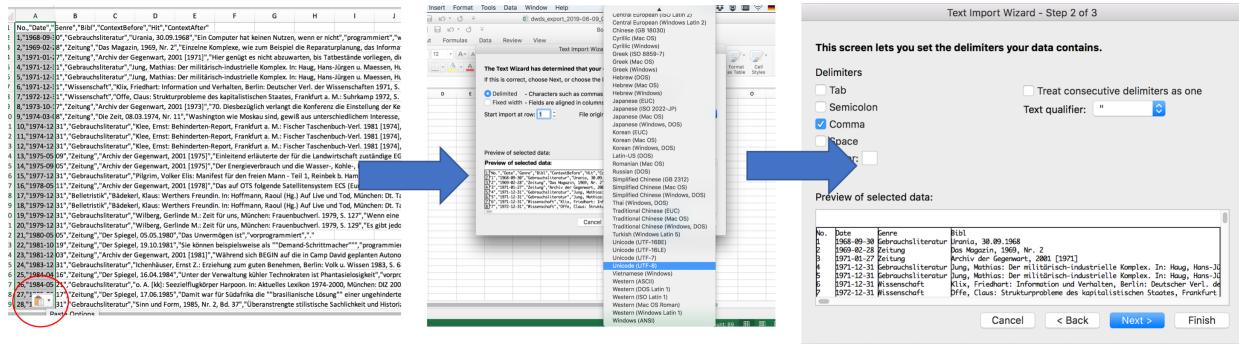


Abbildung 4: Import in Excel

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“) – 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 halten)
7. Dieser letzte Schritt erübriggt 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...) Tipp: Um alle Spalten auf einmal als „Text“ zu formatieren, einfach im Fenster ganz nach rechts scrollen und mit gedrückter Shift-Taste auf die letzte Spalte klicken, dann sind alle Spalten markiert.

2.2.3.2 Import in Calc

Öffnet man die Datei im kostenlosen Tabellenkalkulationsprogramm Calc von LibreOffice (mit Rechtsklick > Öffnen 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.

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.

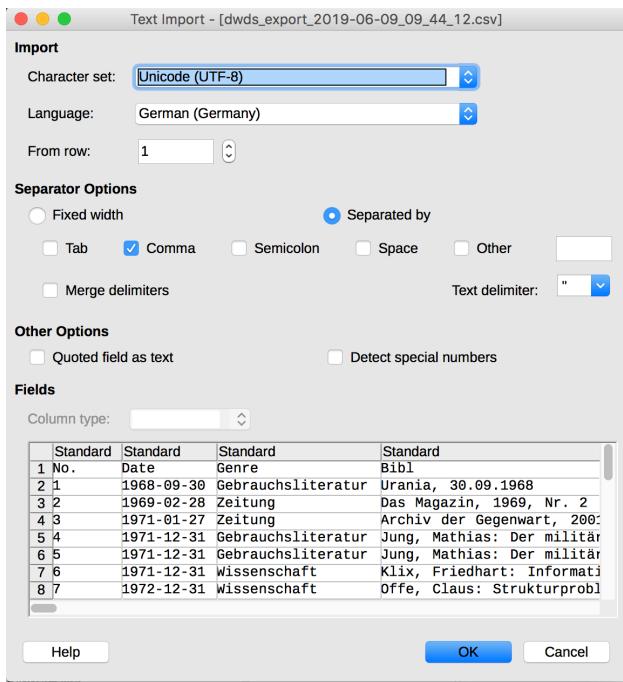


Abbildung 5: Import in Calc

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 prädiktative 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? – Während beispielsweise Computer oder Roboter im wörtlichen Sinne programmiert werden, bezieht sich der Begriff bei Krisen und Katastrophen darauf, dass Voraussetzungen geschaffen wurden, die unausweichlich den thematisierten unschönen Ausgang zur Folge haben. Es liegt also ein metaphorischer Gebrauch vor, bei der Aspekte der Quelldomäne „Technik“ auf eine abstraktere Zieldomäne übertragen werden.

In den nächsten Abschnitten wollen wir uns beiden Fragen etwas genauer zuwenden.

3.1.1 Annotation prädikativ vs. nicht-prädikativ

Wenn wir Daten annotieren, besteht eine wesentliche Herausforderung immer in der **Operationalisierung** konkreter Fragestellungen. In vielen Fällen ist es so, dass wir die Frage, die uns interessiert, auf den ersten Blick für jeden Datenpunkt klar beantworten zu können glauben. Bei genauerem Hinsehen ergeben sich dann aber doch einige Zweifelsfälle. So ist es auch hier: Um die Frage operationalisieren zu können, muss man zunächst einmal die Entscheidung treffen, ob man eine Struktur wie *Der Computer ist programmiert* als Zustandspassiv mit *sein* als Hilfsverb (analog zum Vorgangspassiv mit *werden* als Hilfsverb) oder als

The screenshot shows two parts of an Excel spreadsheet. On the left, a 'Create Table' dialog box is open over a table of news articles. The dialog box has fields for 'Where is the data for your table?' containing '\$A\$1:\$G\$69', 'My table has headers', and buttons for 'Cancel' and 'OK'. Above the dialog, the Excel ribbon is visible with tabs like Home, Insert, Page Layout, Formulas, Data, Review, View, and PivotTable. A red circle highlights the 'Insert' tab icon. On the right, the completed table is shown with an additional column 'praedikativ' at the end of each row. A blue arrow points from the 'Create Table' dialog to the completed table.

Abbildung 6: Formatierung als Tabelle und Hinzufügen einer Annotationsspalte *praedikativ*

Konstruktion aus der Kopula *sein* und dem Partizip II *programmiert* interpretiert. Wir entscheiden uns hier für Letzteres. Jedoch zeigt dieses Beispiel: Wie wir Daten interpretieren, hängt oft genug von unserem theoretischen Zugang ab. Das ist nicht weiter schlimm, sondern liegt in der Natur der Sache – Wissenschaft kann nie ganz frei von Theorie und nie ganz frei von Interpretation sein. Wichtig ist, dass die Entscheidung, die wir treffen, sich gut begründen lässt und konsequent durchgehalten wird.

Wie setzen wir die Annotation nun in unserer Tabelle um? Auch hier zeige ich wieder Wege für Excel und Calc. Gerade die unten skizzierte Möglichkeit, Daten als „Tabelle“ zu formatieren, finde ich persönlich an Excel sehr hilfreich, weshalb ich Excel i.d.R. bevorzuge. Allerdings halte ich es auch für wichtig, sich in der Wissenschaft nicht von proprietärer Software oder proprietären Datenformaten abhängig zu machen, und nicht jede Uni hat eine Office-Lizenz – deshalb zeige ich auch den Weg mit der freien Alternative auf.

3.1.1.1 Umsetzung in Excel

Zunächst empfiehlt es sich, die Tabelle im Excel-Standardformat .xlsx zu speichern.

Excel bietet die schöne Möglichkeit, Daten als Tabelle zu formatieren. Das ist über den Reiter Einfügen > Tabelle möglich, wie in 6 gezeigt. In der Regel erkennt Excel automatisch die Dimensionen der Tabelle, sodass Sie nur noch anklicken müssen, dass die Tabelle Überschriften hat, und dann auf „OK“ klicken können, und schon sind alle Zellen schön formatiert, und vor allem kann man über die kleinen Pfeilsymbole oben die einzelnen Spalten nach bestimmten Werten filtern, was sich im weiteren Verlauf der Arbeit noch als nützlich erweisen kann. (Letzteres erreicht man auch über Daten > Filter, aber mit der Tabellen-Option wird das Ganze optisch noch ein bisschen hübscher, und vor allem muss man keinen neuen Filter setzen, wenn man eine neue Spalte hinzufügt.)

Um die Belege im Kontext besser lesen zu können, empfiehlt es sich, zunächst ein paar Feinjustierungen in der Formatierung vorzunehmen. So können wir Spalten, die wir derzeit nicht benötigen (z.B. alle Spalten mit Metadaten), zunächst ausblenden. (Nicht löschen! Im Zweifelsfall nie Spalten löschen, wer weiß, wozu man sie später noch benötigt...) Außerdem kann es hilfreich sein, den Text in der Spalte mit dem linken Kontext rechtsbündig zu formatieren und die Breite der einzelnen Spalten so anzupassen, dass man den Beleg und ausreichend viel Kontext lesen kann und doch alle derzeit wichtigen Spalten gleichzeitig auf dem Bildschirm zu sehen sind. Wenn Sie die HTML-Version dieses Dokuments lesen, sehen Sie im weiteren Verlauf von Screencast 6 (nach der Formatierung der Daten als Tabelle), wie eine solche Feinjustierung aussehen kann.

Zeilenumbruch innerhalb von Tabellenspalten

In einigen Fällen, in denen man sehr viel Text im linken und rechten Kontext hat und in denen man für die Annotation auch auf den weiteren Kontext angewiesen ist, kann es sinnvoll sein, die Tabelle so zu formatieren, dass innerhalb der Zelle ein Zeilenumbruch vorgenommen wird. Standardmäßig ist die Tabelle so formatiert, dass jede Zelle nur eine Zeile hat, und was über die Zelle hinausgeht, wird nicht angezeigt (ist aber trotzdem

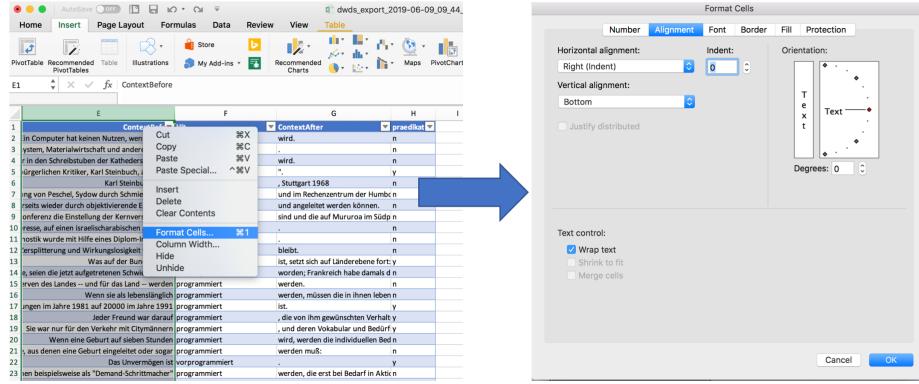


Abbildung 7: Zeilenumbruch innerhalb von Zellen einschalten

Abbildung 8: Eine Tabellenspalte wird so gefiltert, dass nur noch die leeren Zellen zu sehen sind, und allen leeren Zellen wird mit Strg/Cmd+Enter derselbe Wert zugewiesen.

noch in den Daten vorhanden). Wenn man durch Klick auf den Buchstaben oberhalb der Spalte, die man formatieren möchte, zunächst die ganze Spalte markiert, kann man unter Rechtsklick > Zellen formatieren im Tab „Alignment“ („Ausrichtung“) die Option „Wrap text“ (Zeilenumbruch) aktivieren.

Als nächstes fügen wir eine neue Spalte rechts von der letzten existierenden Spalte hinzu, der wir die Überschrift „praedikativ“ geben. (Wir könnten auch problemlos den Umlaut verwenden, aber ich neige dazu, aus Vorsicht alle Sonderzeichen, die Probleme bereiten könnten, wegzulassen.) Hier tragen wir nun für jeden Datenpunkt ein, ob es sich um eine prädiktative Verwendung handelt oder nicht. Ich verwende hierfür gern die Werte „y“ und „n“, weil sie schön kurz sind. j/n oder ja/nein gehen natürlich auch.

Um Zeit zu sparen, kann man auch nur einen der beiden Werte annotieren und dann die leeren Zellen einfach auffüllen, wie in 8 gezeigt: Hier sind die „y“-Werte schon annotiert, alle anderen Zeilen sind leer. Nun filtert man erst die „praedikativ“-Spalte so, dass nur noch die leeren Zellen zu sehen sind, indem man die Zellen mit dem Wert „y“ abwählt. Dann markiert man die Spalte „praedikativ“ von der ersten bis zur letzten Zeile (die Überschrift wird nicht mitmarkiert). Gibt man nun „n“ ein (noch nicht Enter drücken!!), so erscheint der Wert zunächst in der ersten Zeile. Drückt man nun statt der Eingabetaste Strg+Enter (bzw. bei Mac Cmd+Enter), so wird der in der ersten Zeile eingegebene Wert auf alle folgenden Zellen übertragen.

Wenn wir nun den Filter herausnehmen, sehen wir, dass nun alle vorher leeren Zeilen ein „n“ haben, während alle Zeilen mit „y“ unverändert geblieben sind.

Abbildung 9: Formatierung der Tabelle in Calc und Setzen eines Filters

3.1.1.2 Umsetzung in Calc

In Calc empfiehlt es sich, zunächst einmal die Spaltenbreite anzupassen und nicht benötigte Spalten auszublenden (nicht löschen – im Zweifelsfall niemals Spalten oder Zeilen löschen, wer weiß, wofür man sie noch benötigt!). Ich selbst gehe in der Regel so vor, dass ich alle Spalten bis auf diejenigen mit den eigentlichen Belegen (linker Kontext, Treffer, rechter Kontext) ausblende und die Spalte mit dem linken Kontext so formatiere, dass der Text rechtsbündig angezeigt wird. So kann ich bequem den Beleg vom linken Kontext über den Treffer bis zum Keyword lesen. In der HTML-Version dieses Tutorials sehen Sie das in Screencast 9.

Wenn Sie die Formatierungsoptionen für zukünftige Sitzungen speichern möchten, müssen Sie die Datei in einem anderen Format, z.B. im Calc-Standardformat .ods, speichern. Prinzipiell können Sie aber auch einfach in der CSV-Datei weiterarbeiten. Wenn Sie die Datei zwischenspeichern, werden dann eventuell neu eingetragene Daten gespeichert, nicht aber die Formatierung, die Sie dann, wenn Sie die Datei schließen und wieder öffnen, noch einmal neu einstellen müssen.

Wir können nun eine neue Spalte rechts von der letzten existierenden Spalte hinzufügen, der wir die Überschrift „praedikativ“ geben. (Wir könnten auch problemlos den Umlaut verwenden, aber ich neige dazu, aus Vorsicht alle Sonderzeichen, die Probleme bereiten könnten, wegzulassen.) Hier tragen wir nun für jeden Datenpunkt ein, ob es sich um eine prädiktative Verwendung handelt oder nicht. Ich verwende hierfür gern die Werte „y“ und „n“, weil sie schön kurz sind. j/n oder ja/nein gehen natürlich auch.

Um Zeit zu sparen, kann man auch nur einen der beiden Werte annotieren und dann die leeren Zellen einfach auffüllen. Dafür müssen wir zunächst einen Filter setzen, wie in 9 gezeigt. Über diesen Filter können wir jetzt die leeren Zellen ausblenden. Hier sind die „y“-Werte schon annotiert, alle anderen Zeilen sind leer. Nun filtert man erst die „praedikativ“-Spalte so, dass nur noch die leeren Zellen zu sehen sind, indem man die Zellen mit dem Wert „y“ abwählt. Dann markiert man die Spalte „praedikativ“ von der ersten bis zur letzten Zeile (die Überschrift wird nicht mitmarkiert). Gibt man nun „n“ ein (noch nicht Enter drücken!!), so erscheint der Wert zunächst in der ersten Zeile. Drückt man nun statt der Eingabetaste Alt+Enter, so

The figure shows two screenshots of an Excel spreadsheet. On the left, a filter dialog is open over a table. The filter dropdown for column H is set to 'Empty'. The table rows contain German annotations. On the right, the same table is shown with the 'Empty' filter applied, resulting in all empty cells being populated with the letter 'n'. An orange box highlights the 'Alt + Enter' key combination.

Abbildung 10: Eine Tabellenspalte wird so gefiltert, dass nur noch die leeren Zellen zu sehen sind, und allen leeren Zellen wird mit Alt+Enter derselbe Wert zugewiesen.

wird der in der ersten Zeile eingegebene Wert auf alle folgenden Zellen übertragen.

Damit ist die Spalte nun vollständig ausgefüllt.

3.1.2 Annotation metaphorisch vs. nicht-metaphorisch

Für die weitere Annotation können wir die nicht-prädikativen Fälle außer Acht lassen. Hier können wir auf die oben erwähnten Filteroptionen zurückgreifen, um die nicht-prädikativen Fälle herauszufiltern.

Nun gilt es, zu entscheiden, wann *programmiert* und *vorprogrammiert* metaphorisch verwendet werden und wann nicht. Auch das ist auf den ersten Blick denkbar einfach: Einen Computer oder einen Roboter kann man im wörtlichen Sinn programmieren, eine Katastrophe eher nicht – allenfalls indirekt, indem man z.B. Maschinen programmiert, die dann die Weltherrschaft übernehmen, siehe so ziemlich jede Dystopie von „Terminator“ bis „Matrix“. Aber genau dieses indirekte Programmieren bringt uns schon zu möglichen Zweifelsfällen: Was ist, wenn sich ein Satz wie *Die Konfrontation ist programmiert* auf einen Roboter bezieht?

Solche Zweifelsfälle ergeben sich gerade bei einer im weitesten Sinne semantischen Annotation immer. Daher ist es wichtig, klare **Annotationsrichtlinien** zu formulieren, in der alle Annotationsentscheidungen genau dokumentiert sind. Oftmals entwickeln sich diese Richtlinien im Zuge der Annotation selbst, weil man über Daten stolpert, die man so zunächst nicht erwartet hätte. (Was übrigens ein gutes Argument dafür ist, sich bei der Analyse von Sprache nicht allein auf die eigene Intuition zu verlassen, sondern Korpusdaten zu Rate zu ziehen!)

Wenn wir nun wörtlichen und metaphorischen Gebrauch annotieren wollen, könnten unsere Annotationsrichtlinien zunächst ganz einfach so aussehen:

1. Geht aus dem Kontext eindeutig hervor, dass ein Computer bzw. eine Maschine programmiert worden ist, liegt wörtlicher Gebrauch vor.
2. Geht aus dem Kontext eindeutig hervor, dass sich das Verb auf eine andere Entität bezieht, liegt metaphorischer Gebrauch vor.
3. Geht aus dem Kontext nicht hervor, worauf genau sich „(vor)programmiert“ ist, wird der Beleg als unklar gewertet.

Auf diesen Kriterien aufbauend können wir nun eine neue Spalte in unserer Tabelle eröffnen, die wir z.B. „Lesart“ nennen können. Hier vergeben wir die Werte „lit“ (literal/wörtlich), „met“ (metaphorisch) und „unklar“. Gerne können Sie es einmal versuchen und Ihre Ergebnisse dann mit meinen (in den .xlsx- und .ods-Dateien im „data“-Ordner) vergleichen.

Der große Vorteil der oben formulierten Annotationskriterien ist, dass sie sich in den meisten Fällen relativ zweifelsfrei anwenden lassen. Jedoch zeigt sich beim Durchgehen der konkreten Belege, dass die binäre Unterscheidung „wörtlich/metaphorisch“ dem Gebrauch von *(vor)programmiert* möglicherweise nicht ganz gerecht wird. So fallen die folgenden Beispiele alle in die „metaphorische“ Kategorie:

- (1) Der moderne, verbildete Mensch ist nach festen Rhythmen auf das eingeschaltete Gerät programmiert und genußbereit.
- (2) Unsere Gene sind auf Lug und Trug programmiert
- (3) da ist Streit mit den Arbeitgebern programmiert.

Die ersten beiden Beispiele bedienen sich der verbreiteten „Computermetapher“, konzeptualisieren also den menschlichen Geist bzw. die menschlichen Gene als „Computer“. Das ist im letzten Beispiel nicht der Fall: Hier geht es nicht um das Objekt des Programmierungsvorgangs, sondern um das Resultat. Diese Verwendung ist in gewisser Weise also abstrakter. Das ist allerdings eine Dimension, die grundsätzlich von der Dimension der wörtlichen vs. metaphorischen Verwendung unabhängig ist: Angenommen, ich baue mir, wie es verrückte Wissenschaftler in Filmen gerne tun, eine Frühstücksmaschine, die so programmiert ist, dass sie mir morgens um 7 ein Spiegelei brät, und sage: „Das Spiegelei ist für 7 programmiert“, dann ist das zwar eine resultsbezogene, aber keine metaphorische (sondern eher eine metonymische) Verwendung.

Es wäre daher sinnvoll, auch diese Dimension noch zu kodieren.¹ Deshalb fügen wir noch eine weitere Annotationsspalte hinzu, die wir „Referenz“ nennen: Referiert der fragliche Satz auf das, was programmiert wird, oder auf das Resultat der Programmierung?

Auch hierfür formulieren wir wieder Annotationskriterien:

1. Wenn aus dem Kontext eindeutig hervorgeht, dass sich der Satz auf das Objekt des Programmierungsvorgangs bezieht (*der Computer ist programmiert* „jemand (Subj.) hat den Computer (Obj.) programmiert“), wird der Beleg mit „obj“ annotiert.
2. Wenn aus dem Kontext eindeutig hervorgeht, dass sich der Satz auf das Resultat des (ggf. stark metaphorischen) Programmierungsvorgangs bezieht (*das Spiegelei ist programmiert* „jemand hat die Frühstücksmaschine so programmiert, dass sie ein Spiegelei (Resultat) macht“ oder *die Katastrophe ist programmiert* „es wurden Entscheidungen getroffen, die zwangsläufig in eine Katastrophe (Resultat) führen“), so wird der Beleg mit „res“ annotiert.
3. Lässt sich keine eindeutige Entscheidung treffen, bekommt der Beleg den Wert „unklar“.

In den .xlsx- und .odt-Dateien im „data“-Ordner habe ich das in der Spalte „Referenz“ umgesetzt. Auch hier können Sie gern die Probe aufs Exempel machen und überprüfen, ob Ihre Annotationen mit meinen übereinstimmen. Wahrscheinlich werden Sie im einen oder anderen Fall andere Entscheidungen treffen als ich – das ist ganz normal und auch der Grund dafür, warum man idealerweise mindestens zwei Personen unabhängig voneinander annotieren lassen und dann die Annotationen vergleichen sollte. (De facto ist das natürlich gerade bei einer Seminararbeit häufig nicht möglich).

3.2 Auswertung und Visualisierung

Nachdem wir nun die Daten annotiert haben, können wir unsere Annotationen quantitativ auswerten. Auch hier zeige ich wieder die einzelnen Wege für Excel und Calc auf.

¹ Der Vollständigkeit halber sei darauf hingewiesen, dass es sich dabei um eine Post-hoc-Analyse handelt. Wenn Sie sich ein wenig in die Wissenschaftsphilosophie einlesen, werden Sie merken, dass so etwas nicht unumstritten ist: Oft gilt es als Ideal, sämtliche Hypothesen und Analysemethoden im Voraus festzulegen, bevor man sich den Daten selbst zuwendet. *Post hoc* aufgestellte Hypothesen müsste man dann eigentlich anhand von neuen Daten überprüfen. De facto ist es freilich oft so, dass für so ein rigides Vorgehen Zeit und Ressourcen fehlen. Gerade bei einer Seminararbeit können Sie diesen Punkt natürlich in aller Regel getrost ignorieren.

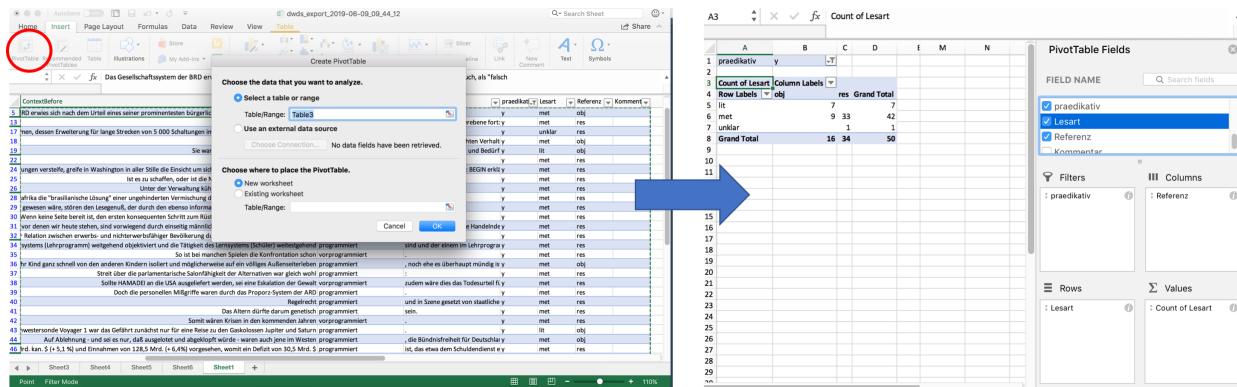


Abbildung 11: Erstellen einer Pivot-Tabelle in Excel.

Row Labels	Count of Hit
programmiert	70
vorprogrammiert	18
Grand Total	88

Abbildung 12: Eine Pivot-Tabelle in Excel.

3.2.1 Auswertung und Visualisierung in Excel

Ideal für die Auswertung und Visualisierung in Excel ist die PivotTable-Funktion. Manches an dieser Funktion ist zunächst ein wenig gewöhnungsbedürftig, aber nach kurzer Eingewöhnungszeit ist sie doch halbwegs logisch und intuitiv.

Stellen Sie zunächst sicher, dass eine Zelle innerhalb der Tabelle angewählt ist (z.B. die Zelle ganz oben links). Jetzt klicken wir im Reiter „Einfügen“ auf „PivotTable“. Nun öffnet sich ein Fenster, in dem wir gefragt werden, welche Zellen Teil der PivotTable werden sollen (hier sollte Excel bereits automatisch erkannt haben, dass wir die ganze Tabelle einbeziehen wollen, sodass wir nichts mehr ändern müssen) und ob die Tabelle auf dem aktuellen oder einem neuen Arbeitsblatt erstellt werden soll – es empfiehlt sich, sie auf einem neuen Arbeitsblatt zu erstellen, was auch die Default-Option ist. Also können wir einfach OK klicken. In der HTML-Version dieses Tutorials können Sie die einzelnen Schritte in Screencast 11 nachverfolgen.

Nun öffnet sich ein neues Arbeitsblatt (mit den Reitern unten können Sie zwischen den Arbeitsblättern navigieren und ihnen ggf. auch aussagekräftigere Namen geben). Wir sehen ein dreigeteiltes Fenster. Im Arbeitsblatt selbst finden wir ein etwas kryptisch aussehendes, noch weitgehend leeres Feld mit einer Beschriftung wie „PivotTable1“ o.ä. Das ist quasi der Platzhalter für die noch zu erstellende Tabelle. Rechts sehen wir oben eine Aufstellung der Namen der Tabellenspalten, unten sehen wir ein wiederum viergeteiltes Fenster. In die vier Felder in diesem Fenster können wir nun ausgewählte Spaltennamen aus dem Fenster oben rechts ziehen. Probieren Sie doch einmal, die Spalte „Hit“ in das Feld „Zeilen“ zu ziehen. Jetzt sehen Sie in der Pivot-Tabelle die beiden Zeilen „programmiert“ und „vorprogrammiert“. Höchstwahrscheinlich ist „Hit“ auch automatisch im Fenster „Werte“ unten rechts aufgetaucht. Deshalb wird Ihnen in der Pivot-Tabelle auch die Häufigkeit der beiden Varianten angezeigt, wie in 12.

Die Logik ist also ganz einfach: Was im Feld „Spalten“ steht, taucht in den Spalten der Tabelle auf, was im Feld „Zeilen“ steht, taucht in den Zeilen auf, und was im Feld „Werte“ steht, das wird ausgezählt². Mit Hilfe des Felds „Filter“ kann man die Daten bei Bedarf filtern.

In unserem Beispiel wollen wir genau das tun: Wir wollen ja nur die prädikativ gebrauchten Instanzen von

²bzw. bei numerischen Werten aufsummiert; hier muss man ggf. aufpassen, dass die richtige Operation gewählt ist. Durch Klick auf das kleine Info-Symbol in den Feldern kann man das bei Bedarf anpassen

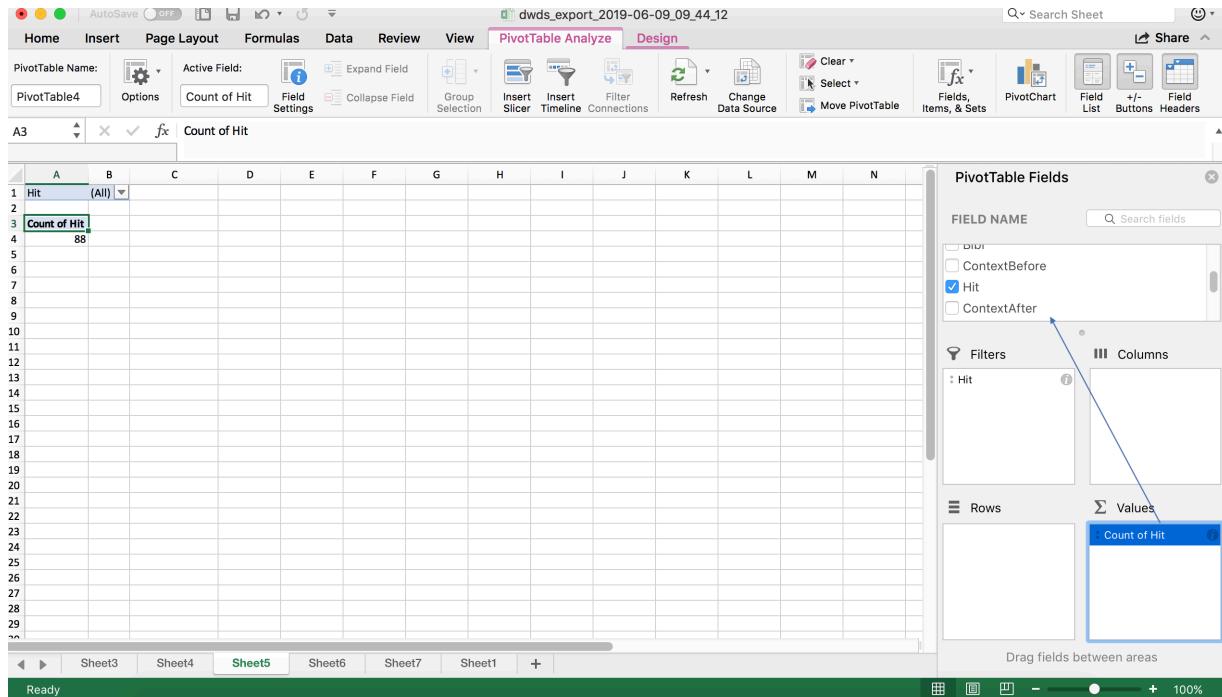


Abbildung 13: Entfernen einer Spalte aus dem „WerteFeld.“

(vor)programmiert berücksichtigen. Leeren wir die beiden Felder zunächst wieder, indem wir „Hit“ aus dem jeweiligen Feld in das Fenster rechts oben ziehen (Fig. 13).

Dann ziehen wir die Spalte „prädikativ“ in das Feld „Filter“. Jetzt können wir im Fenster links die nicht-prädiktiven Daten herausfiltern (s. Fig. 14).

Nun wollen wir eine tabellarische Übersicht über die Lesarten und die Referenz (Objekt vs. Resultat) getrennt nach den beiden Varianten „programmiert“ und „vorprogrammiert“ bekommen. Führen wir uns noch einmal vor Augen, was die dafür relevanten Tabellenspalten sind:

- Die Spalte „Lesart“ enthält die Lesarten.
- Die Spalte „Referenz“ enthält die Information darüber, ob der jeweilige Beleg auf das Objekt des Programmierungsvorgangs oder dessen Resultat referiert.
- Die Spalte „Hit“ enthält die Variante des Treffers.

Zur Auswertung müssen wir die drei Spalten nun sinnvoll auf die „Zeilen“- und „Spalten“-Felder verteilen und zudem angeben, was ausgezählt werden soll. Hier gibt es mehrere Möglichkeiten; eine davon ist in 15 dargestellt: In den Spalten werden die Daten nach Variante („Hit“) ausgewertet, in den Zeilen zum einen nach Lesart, zum anderen nach Referenz. Ausgezählt wird die Spalte „Lesart“ – genauso gut könnten wir aber auch die Spalte „Referenz“ auszählen, die Ergebnisse wären die gleichen, da ja beide Variablen in der Tabelle berücksichtigt sind.

Schon auf den ersten Blick sehen wir eine ungleiche Verteilung der Daten auf die beiden Varianten: *vorprogrammiert* wird in unseren Daten ausschließlich in metaphorischen Kontexten und ausschließlich für Resultate gebraucht. Bei *programmiert* ist der Gebrauch vielfältiger, wenngleich auch hier eine Präferenz für eben diese Merkmalskombination (metaphorischer Kontext/Resultat) deutlich wird.

Quasi als Sahnehäubchen können wir diese Verteilung auch visualisieren, beispielsweise mit einem Balkendiagramm.³ Das geht, indem wir die relevanten Zellen in der PivotTable markieren (also alles, was

³Vgl. jedoch z.B. diese Seite zu Problemen, die Balkendiagramme u.U. mit sich bringen.

Abbildung 14: Herausfiltern der Daten mit `praedikativ=n`.

	B	C	D	E	F	G	H	I	J	K	L	M
1	praedikativ	y										
2												
3	Count of Lesart	Column Labels										
4	Row Labels	programmiert	vorprogrammiert	Grand Total								
5	lit	7		7								
6	obj	7		7								
7	met	28		14	42							
8	obj	9		9								
9	res	19		14	33							
10	unklar	1		1	1							
11	res	1		1								
12	Grand Total	36		14	50							
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27												
28												
29												

Abbildung 15: Tabellarische Auswertung der Korpusdaten mit Hilfe der PivotTable-Funktion.

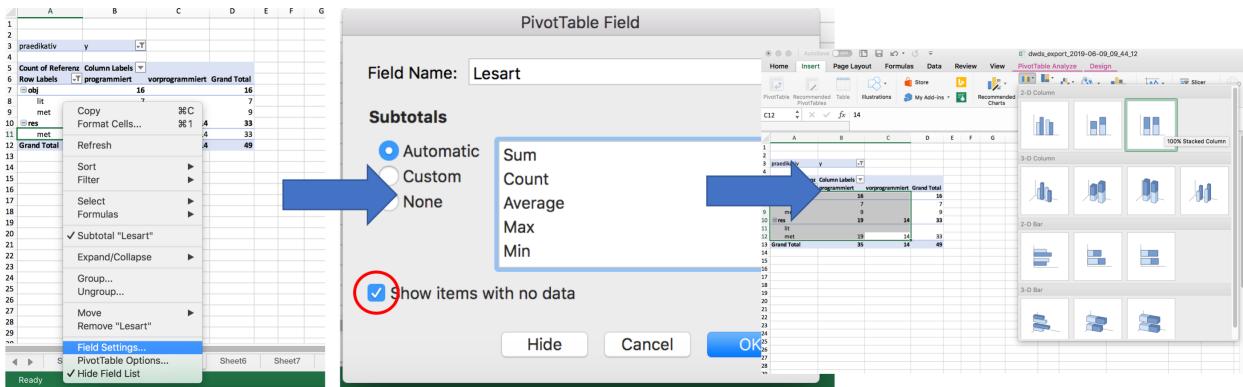


Abbildung 16: Erstellen eines Balkendiagramms aus einer Pivot-Tabelle in Excel.

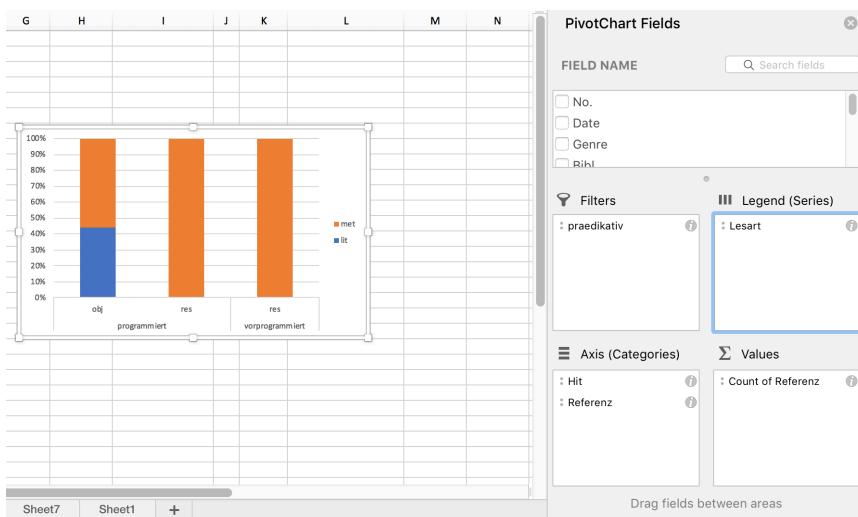


Abbildung 17: Der Dialog 'Pivot Chart Fields' (rechts) erlaubt es, anzupassen, was wo in der Grafik dargestellt wird.

nicht die Gesamtsumme anzeigt) und dann in „Einfügen“ eine passende Visualisierungsoption auswählen. Abbildung 16 zeigt eine Möglichkeit, das umzusetzen:

Welche Daten auf der x- und auf der y-Achse dargestellt und welche innerhalb der Balken farblich kodiert werden, hängt davon ab, welche Daten in der Pivot-Tabelle in den Zeilen und in den Spalten stehen. Gegebenenfalls kann man hier noch ein paar Änderungen vornehmen, um die Darstellung sinnvoller zu gestalten. So entsteht in Fig. 16 ein Diagramm, bei dem die Variante (*programmiert* vs. *vorprogrammiert*) farblich kodiert wird. Das ist nicht unbedingt sinnvoll, weil wir ja wissen wollen, wie sich die Verteilung der unterschiedlichen Lesarten zwischen den beiden Varianten unterscheidet. Man kann die visuelle Darstellung auch in dem „Pivot Chart Fields“-Feld ändern, das sich bei der Erstellung des Balkendiagramms geöffnet haben dürfte (17).

Und noch etwas Feinjustierung: Damit die Grafik nicht so asymmetrisch aussieht wie Fig. 18, werden (im ersten Schritt in 16) zunächst die Feldeinstellungen so verändert, dass auch Felder mit Null-Werten angezeigt werden. Anschließend wird ein Balkendiagramm ausgewählt, das den prozentualen Anteil der jeweiligen Variante darstellt. Durch Rechtsklick auf die Balken kann man außerdem noch „Data Labels“ hinzufügen, d.h. die absoluten Werte in den Balken darstellen lassen. Das ist empfehlenswert, weil so die Leserin oder der Leser schnell einen Eindruck gewinnen kann, wie groß die Datenbasis ist, auf der die Darstellung basiert

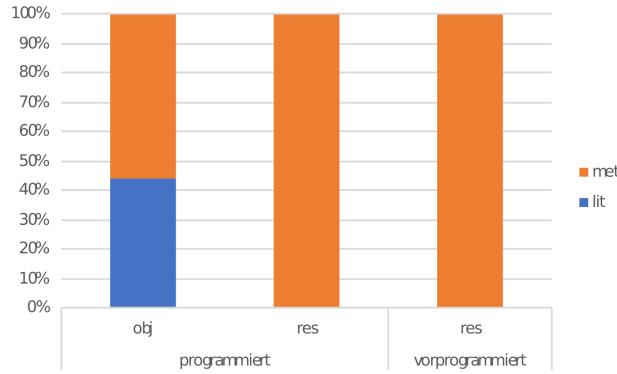


Abbildung 18: Balkendiagramm, in dem nicht belegte Variablenausprägungen ausgelassen werden.

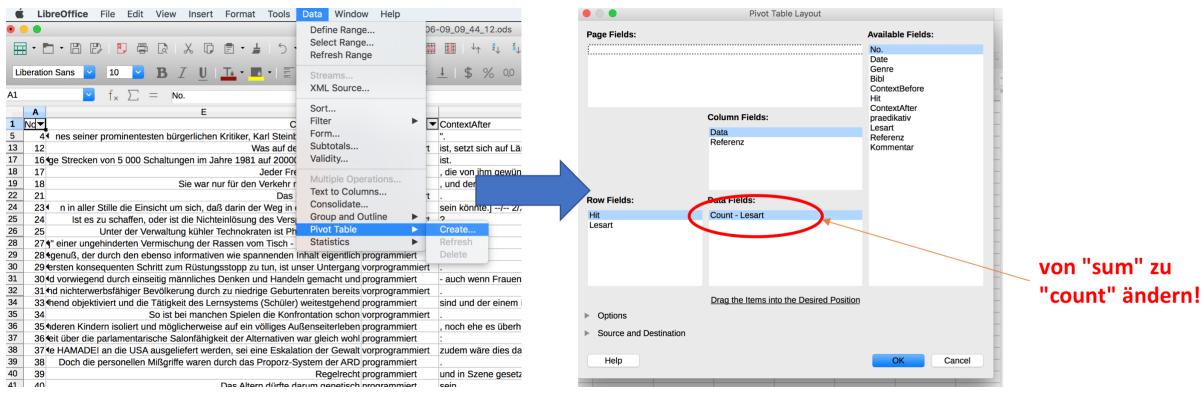


Abbildung 19: Erstellen einer Pivot-Tabelle in Calc.

– denn es macht ja schon einen gewichtigen Unterschied, ob eine Verteilung wie, sagen wir 10% : 90% auf zehn, auf hundert oder auf tausend Datenpunkten basiert! Die „Data Labels“ kann man hinzufügen durch Rechtsklick auf die Balken und Klick auf den Punkt „Show data labels“.

3.2.2 Auswertung und Visualisierung in Calc

Auch in Calc gibt es eine PivotTable-Funktion, die sich im Dropdown-Menü „Data“ findet. Mit Hilfe von PivotTable > Erstellen gelangt man in ein Auswahlmenü, in dem rechts die einzelnen Spaltennamen als „available fields“ zu sehen sind. Links sehen wir drei rechteckige Felder: eins für Zeilen, eins für Spalten und eins für Werte (das Letztere trägt die nicht ganz so aussagekräftige Überschrift „Data fields“). Aus der Auflistung rechts können wir nun die Variablen, die wir in den Spalten sehen wollen, ins „Spalten“-Feld ziehen, diejenigen, die wir in den Zeilen sehen wollen, ins „Zeilen“-Feld und schließlich diejenigen, die Calc für uns auszählen soll, ins „Data fields“-Feld.

Beim Erstellen der Pivot-Tabelle können wir in den ausklappbaren Optionen auch auswählen, dass wir keine Zeilen- und Spaltensummen sehen wollen (die brauchen wir nicht, zumal sie dann ggf. auch in den auf der Tabelle basierenden Visualisierungen dargestellt werden und da nur verwirren), aber dass wir Filter setzen möchten. Wenn wir nun auf OK klicken, sehen wir die Pivot-Tabelle, oberhalb derer sich ein „Filter“-Feld befindet, auf das wir doppelklicken können, um durch Auswählen des Attribut-Wert-Paars „praedikativ = y“ die Tabelle so zu filtern, dass nur die prädiktiven Belege angezeigt werden.

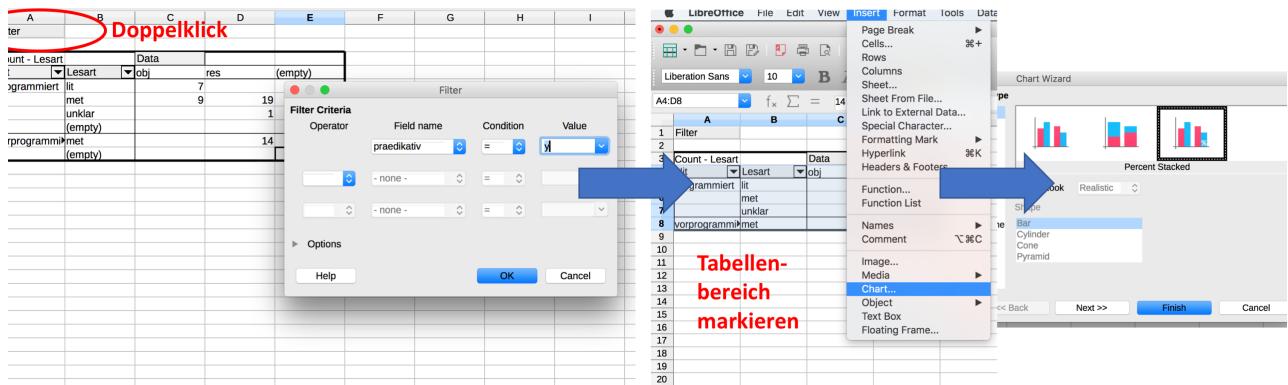


Abbildung 20: Erstellen einer Pivot-Tabelle in Calc.

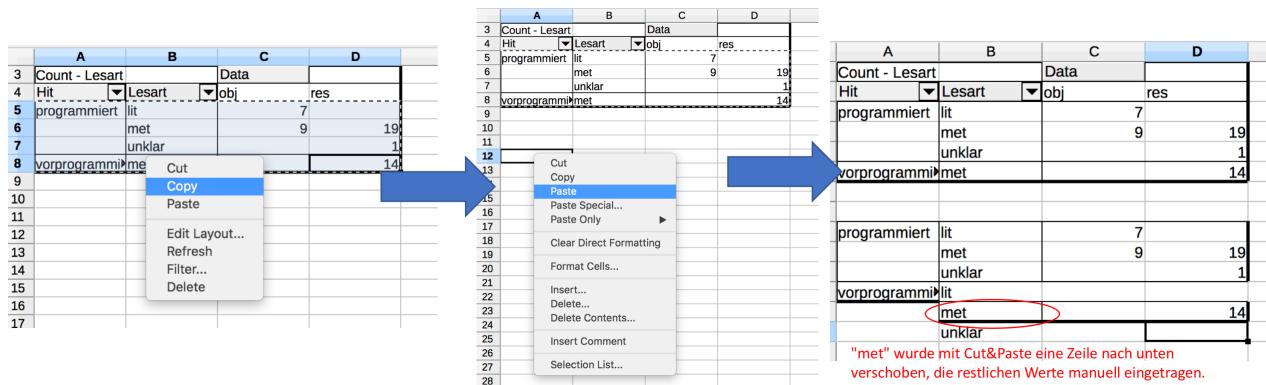


Abbildung 21: Erstellen einer Pivot-Tabelle mit Nullwerten in Calc (Workaround).

Mit Hilfe der Tabelle können wir nun die Verteilung auch visualisieren, z.B. mit einem Balkendiagramm.⁴ Das ist über Insert > Chart möglich. Hier haben wir die Wahl zwischen mehreren Optionen und wählen ein sog. gestapeltes Balkendiagramm, das die prozentuale Verteilung anzeigt. Die einzelnen Schritte sind in 20 dargestellt.

Im Gegensatz zu Excel bietet Calc derzeit leider keine Möglichkeit, sich in Pivot-Tabellen Nullwerte anzeigen zu lassen. Das heißt, wenn wir im Balkendiagramm die Kategorien „lit“ und „unklar“ auch für „vorprogrammiert“ sehen wollen, um das Diagramm symmetrisch zu halten, müssen wir sie irgendwie manuell einfügen, weil für „vorprogrammiert“ ja nur metaphorische Lesarten belegt sind. Einen sehr uneleganten, aber funktionierenden Workaround zeigt 21: Weil wir an der Pivot-Tabelle selbst nichts verändern können, copy&pasten wir sie an eine andere Stelle im Arbeitsblatt und ergänzen die fehlenden Kategorien manuell. Das ist zwar, wie es ein User in einem Frage-und-Antwort-Forum zu diesem Thema sehr schön formuliert hat, weniger eine Lösung als eine Kapitulation, aber zumindest für diesen recht überschaubaren Datensatz funktioniert der Ansatz.

Zuletzt wollen wir die Balken noch mit Labels versehen, d.h. wir wollen die absoluten Werte auf den Balken anzeigen. Auch hier kann Excel etwas mehr als Calc, und wir müssen etwas tricken, um die absoluten Werte anzeigen zu lassen. Mit Rechtsklick auf einen der Balken > Insert Data Labels können wir zunächst die Labels anzeigen lassen, sehen aber erstens keine absoluten Werte, sondern Prozentwerte, und zweitens absurde Prozentwerte, weil Calc offenbar die Zahlen in der Tabelle für relative Werte hält. Mit Doppelklick auf eines der Labels können wir jedoch das Zahlenformat ändern und angeben, dass wir die Labels einfach als „Text“ angezeigt bekommen wollen – also einfach das, was in der Pivot-Tabelle selbst steht (s. Fig. 22).

⁴Vgl. jedoch den Caveat in der vorherigen Fußnote.

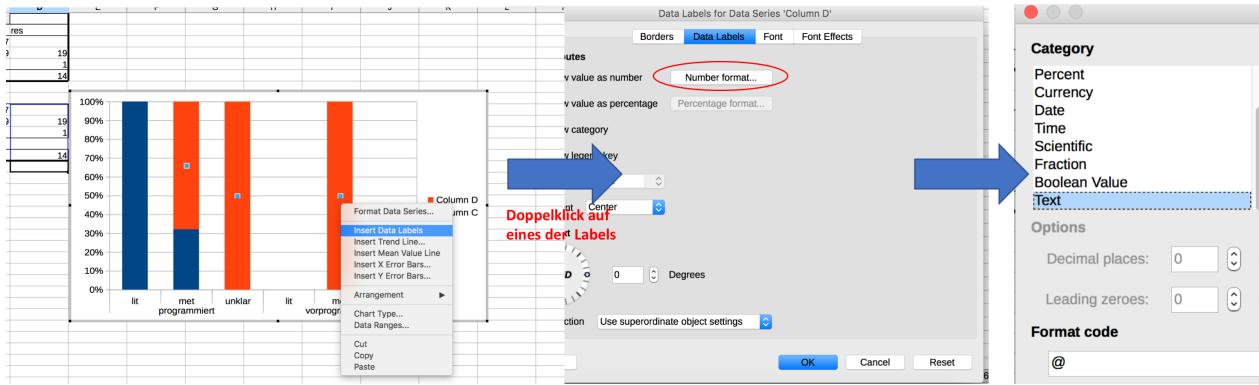


Abbildung 22: Einfügen von Data labels in Calc.

4 Und nun...?

Fassen wir kurz die wesentlichen Schritte zusammen, die wir uns in diesem Tutorial näher angeschaut haben:

- Zunächst haben wir eine Fragestellung formuliert und sie korpuslinguistisch operationalisiert.
- Wir haben Daten erhoben, indem wir im DWDS-Kernkorpus nach den beiden Wortformen, die uns interessieren, gesucht haben. Wir haben diese Daten exportiert und haben gesehen, wie man sie in ein Tabellenkalkulationsprogramm einliest.
- Anschließend haben wir die Daten mit grammatischen und semantischen Annotationen versehen.
- Zuletzt haben wir die Daten ausgewertet und visualisiert.

Das allein macht natürlich noch keine gute linguistische Studie aus. Wir müssen uns jetzt überlegen, wie die Daten zu interpretieren sind. Nicht zuletzt müssen wir aber auch unser Vorgehen kritisch hinterfragen und überlegen, welche Unzulänglichkeiten es möglicherweise mit sich bringt.

4.1 Interpretation und Einordnung

Fangen wir mit dem ersten Punkt an, der Interpretation: In unseren Daten wird *vorprogrammiert* ausschließlich im metaphorischen Sinn verwendet und nur dann, wenn auf Resultate Bezug genommen wird. Das untermauert unsere Annahme, dass *vorprogrammiert* gegenüber *programmiert* eine bestimmte semantische Nische einnimmt (wenngleich es in dieser Nische durchaus stark mit der Variante ohne die Partikel *vor-* konkurriert).

Auch werden Sie sich gefragt haben, ob denn das hier durchgearbeitete Beispiel wirklich Stoff für z.B. eine ganze Seminararbeit bietet, wo wir doch am Ende der langwierigen Korpusrecherche nicht viel mehr bekommen haben als ein relativ simples Balkendiagramm. (Formulieren Sie diesen Satz gerne etwas griffiger, damit Sie ihn sich auf ein T-Shirt drucken lassen können.)

Auch wenn das hier besprochene Beispiel durchaus noch ausbaufähig ist (siehe Methodenkritik), bietet es meines Erachtens doch genug Stoff für eine gute Hausarbeit, sofern man die Korpusanalyse in eine gute theoretische Diskussion einbettet und die hier vorgenommene quantitative Analyse vielleicht noch durch die qualitative Analyse von Einzelbelegen ergänzt. Das könnte zum Beispiel so aussehen, dass man sich zunächst in einem Theorienteil mit der semantischen Entwicklung von (nahe-)synonymen Wörtern befasst und dann am Beispiel von (*vor*)*programmiert* der Frage nachgeht, ob die Varianten mit und ohne Partikel unterschiedlich gebraucht werden, also unterschiedliche semantische „Nischen“ ausfüllen. In einem Methodenteil kann man dann die Annotationskriterien offenlegen, ggf. Problemfälle schildern und aufzeigen, wie sie gelöst wurden. Hier kann man auch schon die Grenzen des gewählten Vorgehens ansprechen, auf die wir im nächsten Abschnitt noch näher eingehen werden. Es folgen die quantitative Analyse und die Diskussion der Ergebnisse

vor dem Hintergrund der Forschungsfrage, die gerne mit der qualitativen Analyse von Korpusbelegen gespickt sein darf. In einem Schlussteil werden dann die Ergebnisse zusammengefasst, und es werden Desiderata für zukünftige Forschungen aufgezeigt – ein Punkt, auf den wir ebenfalls im nächsten Abschnitt noch eingehen werden.

4.2 Methodenkritik und offene Fragen

Das führt uns zur Methodenkritik: Wie belastbar sind unsere Ergebnisse? Hier ist als möglicherweise problematischer Punkt zunächst die Stichprobengröße zu nennen. Im DWDS-Kernkorpus haben wir gerade einmal 88 Treffer gefunden, von denen nur rund 50 die prädiktative Verwendung instanziieren, die uns interessiert. Insofern wäre zu fragen, ob wir möglicherweise besser ein anderes Korpus wählen sollten.

Zur Stichprobengröße

Zur Frage nach der Stichprobengröße zitiere ich mich ausnahmsweise mal selbst:

Die wahrscheinlich am häufigsten gestellte Frage von Studierenden, die zum ersten Mal korpuslinguistisch arbeiten, ist: „Wie groß muss meine Stichprobe sein?“ Darauf gibt es leider keine pauschale Antwort. Es gibt keine feste Untergrenze, ab der eine Stichprobe repräsentativ ist (zumal es „echte“ Repräsentativität in dem Sinne, dass die Stichprobe ein ganz genaues Abbild der Grundgesamtheit, nur eben im Kleinen, darstellt, ohnehin nicht geben kann). Die Wahl der Stichprobengröße ist also von mehreren ganz praktischen Faktoren abhängig, unter anderem:

- a) Wie werden die Daten annotiert? Sehr viele Annotationen, die noch dazu erfordern, dass der Kontext mit einbezogen wird, sind zeitaufwendig und rechtfertigen eine kleinere Stichprobe. Arbeitet man dagegen nur mit den reinen Type- und Tokenfrequenzen, ohne eigene Annotationen hinzuzufügen, gibt es keinen Grund, überhaupt eine Stichprobe zu nehmen. In diesem Fall kann man gleich alle Daten mit einbeziehen.
- b) Wie werden die Daten ausgewertet? In manchen Fällen kann man schon mit 100 Belegen aussagekräftige Ergebnisse erzielen. Aber wenn man ein Korpus diachron auswerten möchte, das in 10 Zeitschnitte unterteilt ist, sind 100 Belege offensichtlich zu wenig – denn dann hat man bei gleicher Verteilung gerade einmal 10 Belege pro Zeitschnitt!

Für die ersten Gehversuche z.B. in Seminararbeiten empfehle ich in der Regel, mit 100 bis 500 Belegen zu arbeiten. In den meisten Fällen genügt das, um Tendenzen aufzuzeigen, und ist vom Arbeitsaufwand her auch für AnfängerInnen bewältigbar. Aus den obigen Überlegungen sollte jedoch klar geworden sein, dass diese Zahlen völlig willkürlich sind.

— aus: Hartmann, Stefan. 2018. Deutsche Sprachgeschichte. Grundzüge und Methoden. Tübingen: Francke, S. 206

Weiterhin müssen wir bedenken, dass wir die Verteilung der Daten nur relativ oberflächlich ausgewertet haben. Streng genommen müssten wir eine Reihe von möglicherweise problematischen Aspekten zusätzlich bedenken, denn bei Korpusdaten gibt es immer viele Faktoren, die zu unerwünschten Verzerrungen führen können. So können viele verschiedene Belege von der gleichen Person oder gar aus dem gleichen Text stammen. Tab. 1 zeigt, dass 16 Belege in unserer Konkordanz, wenig überraschend, aus dem „Lexikon der Informatik“ stammen... In der modernen Korpuslinguistik tendiert man daher dazu, statistische Methoden zu verwenden, mit denen man solche Variablen mit einbeziehen kann.

Das Stichwort „statistische Methoden“ bringt uns zu einem weiteren Punkt, den wir hier explizit **nicht** behandelt haben: Wir haben uns nicht mit der Frage beschäftigt, ob der Unterschied zwischen den untersuchten Variablen in irgendeiner Weise statistisch signifikant ist (und was statistische Signifikanz überhaupt bedeutet). In unserem Beispiel war der Unterschied zwischen *programmiert* und *vorprogrammiert* erfreulicherweise so klar und offensichtlich, dass es einer statistischen Analyse nicht wirklich bedarf. Wenn Sie sich weiter damit beschäftigen wollen, empfehle ich die Lektüre einer der einschlägigen Einführungen.

Zum Weiterlesen: Statistik für LinguistInnen

Tabelle 1: Texte in der Konkordanz, die mehr als einen Beleg für (vor)programmiert enthalten

Text	Belege
Rechenberg, Peter: Was ist Informatik?, München: Hanser 1994 [1991]	16
Klee, Ernst: Behinderten-Report, Frankfurt a. M.: Fischer Taschenbuch-Verl. 1981 [1974]	3
Alt, Franz: Liebe ist möglich, München: Piper 1985	2
Archiv der Gegenwart, 2001 [1975]	2
Bädeker, Klaus: Werthers Freundin. In: Hoffmann, Raoul (Hg.) Auf Live und Tod, München: Dt. Taschenbuch-Verl. 1983 [1979]	2
Die Zeit, 22.07.1999, Nr. 30	2
Jung, Mathias: Der militärisch-industrielle Komplex. In: Haug, Hans-Jürgen u. Maessen, Hubert (Hgg.) Kriegsdienstverweigerer - Gegen die Militarisierung der Gesellschaft, Frankfurt a. M.: Fischer 1971	2
Ketman, Per u. Wissmach, Andreas: DDR - ein Reisebuch in den Alltag, Reinbek bei Hamburg: Rowohlt 1986	2
Loriot [d.i. Vicco von Bülow]: Sehr verehrte Damen und Herren, Zürich: Diogenes 1993	2
Wilberg, Gerlinde M.: Zeit für uns, München: Frauenbuchverl. 1979	2

Hier eine Auswahl an deutsch- und englischsprachigen Einführungswerken in die Statistik, die sich explizit an Linguist*innen richten (chronologisch geordnet):

- Vanhove, Jan. 2018. Statistische Grundlagen. Eine Einführung mit Beispielen aus der Sprachforschung. <https://homeweb.unifr.ch/VanhoveJ/Pub/Statistikkurs2/StatistischeGrundlagen.pdf>. (zuletzt abgerufen am 14.06.2019)
- Desagulier, Guillaume. 2017. Corpus linguistics and statistics with R: introduction to quantitative methods in linguistics. New York, NY: Springer.
- Levshina, Natalia. 2015. How to do linguistics with R. Data exploration and statistical analysis. Amsterdam, Philadelphia: John Benjamins.
- Gries, Stefan Th. 2013. Statistics for Linguistics with R: A Practical Introduction. 2nd ed. Berlin, New York: De Gruyter.
- Meindl, Claudia. 2011. Methodik für Linguisten: Eine Einführung in Statistik und Versuchsplanung. Tübingen: Narr.
- Baayen, R. H. 2008. Analyzing Linguistic Data. A Practical Introduction to Statistics using R. Cambridge: Cambridge University Press.
- Butler, Christopher. 1985. Statistics in Linguistics. Oxford: Blackwell.

Über die Methodenkritik hinaus können wir natürlich auch Desiderata formulieren, also fragen, in welche Richtung wir weiterforschen können. So haben wir uns in unserem Beispiel auf prädiktative Verwendungen von (vor)programmiert beschränkt. Das war eine sinnvolle und gut begründbare Vorentscheidung, um den Skopus der Untersuchung einzuschränken, und deshalb nichts, was wir im Rahmen der Methodenkritik hinterfragen sollten. Gleichwohl können wir die Frage stellen, wie sich wohl die flektierten Formen von (vor)programmieren verhalten und ob sich ähnliche Tendenzen auch hier aufzeigen lassen.

Es bietet sich immer an, eine Arbeit mit solchen Desiderata abzuschließen, denn keine Forschungsfrage ist je vollständig beantwortet, und jede (Teil-)Antwort wirft nahezu zwangsläufig neue Fragen auf.