

Einfache Korpusanalysen: Ein Schnelleinstieg

Stefan Hartmann

2019-06-10

Inhalt

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	3
2.2.3 Import in ein Tabellenkalkulationsprogramm	5
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	7
3.1.1.2 Umsetzung in LibreOffice Calc	9
3.1.2 Annotation metaphorisch vs. nicht-metaphorisch	10
3.2 Auswertung und Visualisierung	12
3.2.1 Auswertung und Visualisierung in Excel	12

1 Einstieg

Ziel dieses Tutorials ist es, Anfänger*innen einen möglichst niedrigschwälligen 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.

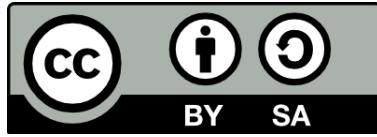
Um wirklich einen Schnelleinstieg bieten zu können, muss ich notwendigerweise vieles vereinfachen. Für Ihre konkrete Korpusstudie werden Sie daher wahrscheinlich nicht umhinkommen, sich an der einen oder anderen Stelle tiefer einzulesen. Dafür verweise ich im Text immer wieder auf weiterführende Ressourcen.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



Dieses Tutorial ist lizenziert unter CC-BY-SA und kann gerne mit Quellenangabe weitergegeben und adaptiert werden.

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 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 Beispielenalyse 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 kompositionally 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 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 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.

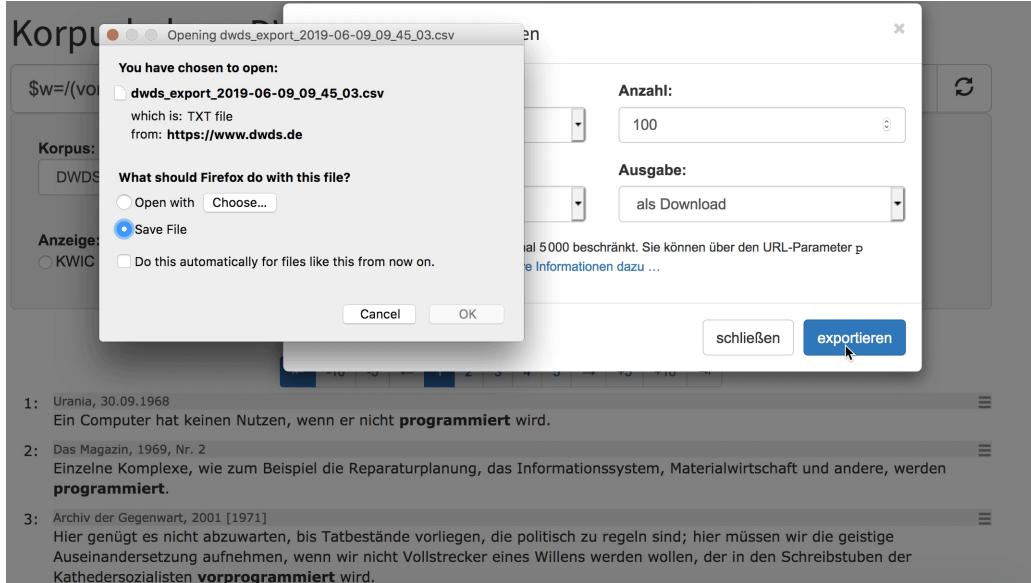


Fig. 1: Export aus dem DWDS

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 r
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","Wissenschaft","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. 98.",Was bedeutet
9 "8","1973-10-17","Zeitung","Archiv der Gegenwart, 2001 [1973]",70. Diesbezüglich verlangt die Konferenz die Einstellung der Kernversuche, die
10 "9","1974-03-09","Zeitung","Die Zeit, 08.03.1974, Nr. 11",Washington wie Moskau sind, gewiß aus unterschiedlichem 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 al
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 Land
16 "15","1977-12-31","Gebrauchsleiteratur","Pilgrim, Volker Eli: 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 Satellitenystem ECS (European Communications Satellites
18 "17","1979-12-31","Belletristik","Bädekerl, Klaus: Werthers Freundin. In: Hoffmann, Raoul (Hg.) Auf Live und Tod, München: Dt. Taschenbuch-Ver
19 "18","1979-12-31","Belletristik","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 beispielsweise 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 so sch
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. (kk): Seesielflugkörper 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 ""brasilianische Lösung"" einer ungehinderten Vermischung di
29 "28","1985-12-31","Gebrauchsleiteratur","Sinn und Form, 1985, Nr. 2, Bd. 37",Überanstrengte stilistische Sachlichkeit und Historizität sowie al
30 "29","1985-12-31","Gebrauchsleiteratur","Ait, Franz: Liebe ist möglich, München: Piper 1985, S. 129",Wenn keine Seite bereit ist, den ersten K
31 "30","1985-12-31","Gebrauchsleiteratur","Ait, 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

```

Fig. 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 fVor SVDafrika die "brasilar programmie.							
29	28	31/12/1985	Gebräuchslit	Sinn und For/überanstrenge stilistische Sachlichkeit programmie ist.							
30	29	31/12/1985	Gebräuchslit	Alt, Franz: Lir Wenn keine Seite bereit ist, den ersten vorprogramm.							
31	30	31/12/1985	Gebräuchslit	Alt, Franz: Lir Die Katastrophen, vor denen wir heute programmie? #NAME?							
32	31	31/12/1985	Gebräuchslit	Zimmerman Zum anderen wird eine fortdauernde Divorprogramm.							
33	32	31/12/1985	Gebräuchslit	Zimmerman FVor die verschiedenen Gebiete des ML programmie (vgl. fVor die gegenwärtig gyältige Planperiode:							
34	33	31/12/1985	Gebräuchslit	Zimmerman Als programmiert Unterricht (PU) wi programmie sind und der einem im Lehrprogramm gespeicherten Lehralgorithmus folgt, der die							
35	34	31/12/1986	Gebräuchslit	Ketman, Pei So ist bei manchen Spielen die Konfront vorprogramm.							
36	35	31/12/1986	Gebräuchslit	Ketman, Pei Als zukünftige Pfarrerin mVor Miria programmie, noch ehe es Vberhaupt mVndig ist: ein Dilemma, das keine einfache Lösung ke							
37	36	23/02/1987	Zeitung	Der Spiegel, Streit Vber die parlamentarische Salon programmie:							
38	37	27/02/1987	Zeitung	Archiv der G Solite HAMADEI an die USA ausgeliefert vorprogramm zudem wVre dies das Todesurteil fVor den EntfVhrten.							
39	38	05/10/1987	Zeitung	Der Spiegel, Doch die personellen Mivgriffe waren programmie.							
40	39	12/09/1988	Zeitung	Der Spiegel, Regelrecht programmie und in Szenen gesetzten von staatlichen Instanzen war vor dem 9. November 1998 kein							
41	40	31/12/1988	Wissenschaft	Weizsäcker Das Altern dVrfte darum genetisch programmie sein.							
42	41	07/04/1989	Zeitung	Archiv der G Somit wVren Krisen in den kommenden vorprogramm.							
43	42	28/08/1989	Zeitung	Der Spiegel, Wie die Schwesterlands Voyager 1 war programmie.							
44	43	31/12/1989	Gebräuchslit	Brandt, Willy Auf Ablehnung - und sei es nur, daß au programmie, die BVndnisfreiheit fVor Deutschland keinesfalls in Erwägung ziehen wollten.							
45	44	31/12/1989	Wissenschaft o. A.: Lexikon-Literatur sowie im allg. Sprachgebrauch vorprogramm.								
46	45	26/02/1991	Zeitung	Archiv der G FVor das Fiskaljahr 1991 sind Ausgaben programmie ist, das etwa dem Schuldendienst entspricht.							
47	46	26/09/1991	Gebräuchslit o. A. [ley]: El Das System kann so programmie werden, daß es beispielsweise Alarm gibt, wenn sich das Schiff einem Hindernis au								

Fig. 3: Konkordanz bei direktem Öffnen in Excel

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 7, sind in 4 visualisiert.
4. 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.
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") - 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 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.

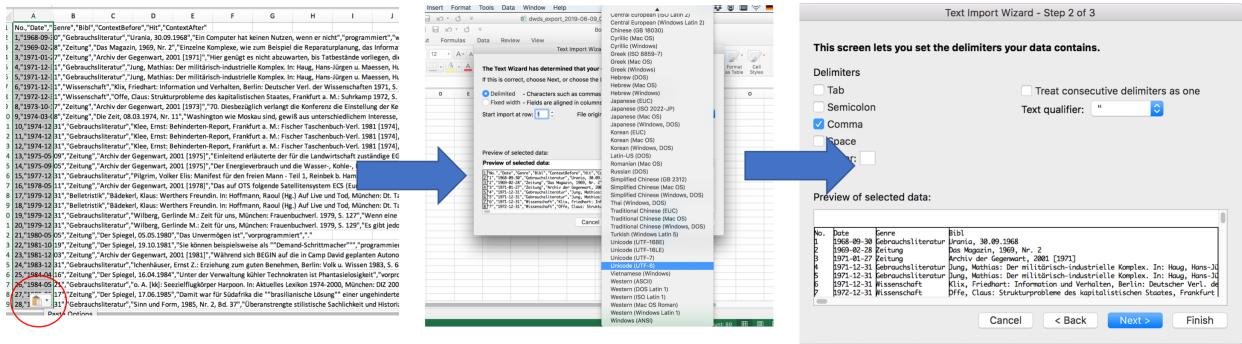


Fig. 4: Import in Excel

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.

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.

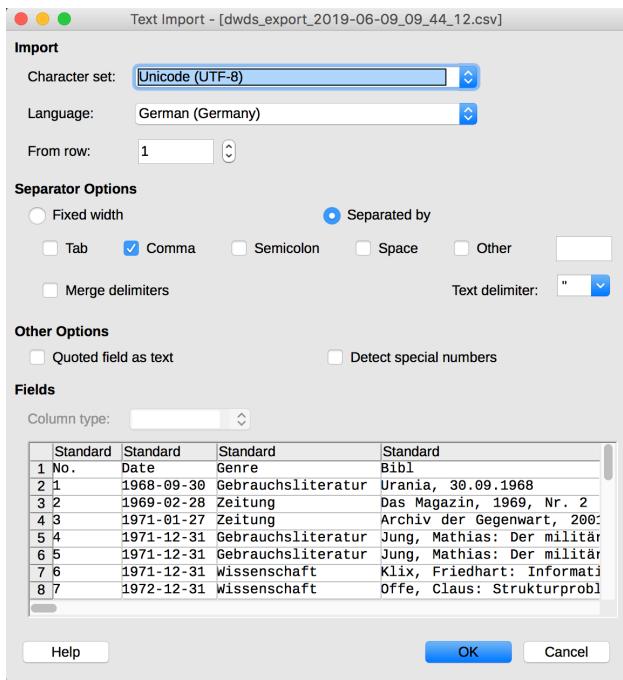


Fig. 5: Import in Calc

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 glauben für jeden Datenpunkt klar beantworten zu können. 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 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,

The screenshot shows the 'Create Table' dialog box in Excel. The range selected is \$A\$1:\$G\$17. The 'My table has headers' checkbox is checked. The 'OK' button is highlighted. To the right is the resulting table in the spreadsheet, which includes an additional column 'praedikativ' at the end of each row.

Fig. 6: Formatierung als Tabelle und Hinzufügen einer Annotationsspalte "praedikativ"

The screenshot shows the 'Format Cells' dialog box for cell E1. The 'Text' tab is selected. Under 'Horizontal alignment', 'Right' is chosen. Under 'Vertical alignment', 'Bottom' is chosen. Under 'Text control', 'Wrap text' is checked. The 'OK' button is visible at the bottom right.

Fig. 7: Zeilenumbruch innerhalb von Zellen einschalten

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 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.

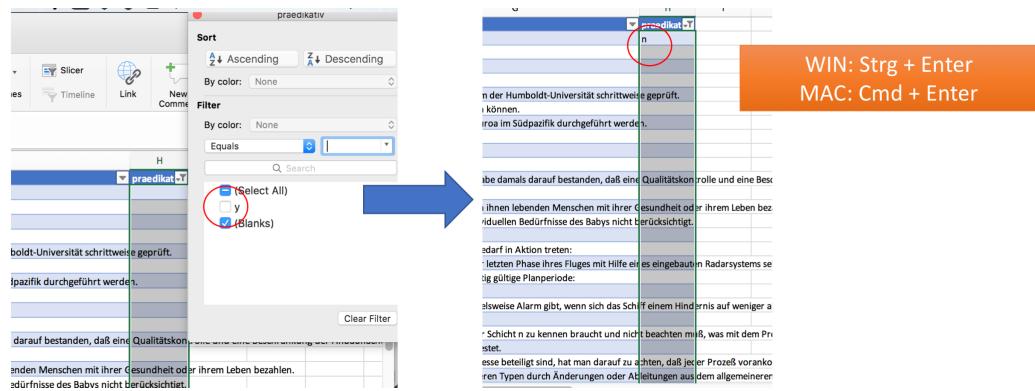


Fig. 8: Eine Tabellenspalte wird so gefiltert, dass nur noch die leeren Zellen zu sehen sind, und allen leeren Zellen wird mit Strg/Cmd+Enter derselbe Wert zugewiesen.

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 prädiktive 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.

3.1.1.2 Umsetzung in LibreOffice 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 prädiktive 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

Fig. 9: Formatierung der Tabelle in Calc und Setzen eines Filters

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 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-prädiktiven Fälle außer Acht lassen. Hier können wir auf die oben erwähnten Filteroptionen zurückgreifen, um die nicht-prädiktiven 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!)

The figure consists of two screenshots of Microsoft Excel. On the left, a filter dialog is open over a table. The dialog shows column H with the dropdown menu 'praedikativ' selected. Under 'Sort Ascending' and 'Sort Descending', there are options 'Top 10', 'Empty', and 'Not Empty'. A checkbox labeled '(empty)' is checked. A blue arrow points from this dialog to the right screenshot. The right screenshot shows the same table after applying the filter. Only the row containing the word 'Empty' in column H is visible. An orange box highlights the text 'Alt + Enter' in the top right corner of the table area.

Fig. 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.

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

¹Der 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*

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 Programmiervorgangs 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) Programmiervorgangs 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.

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.

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.

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.