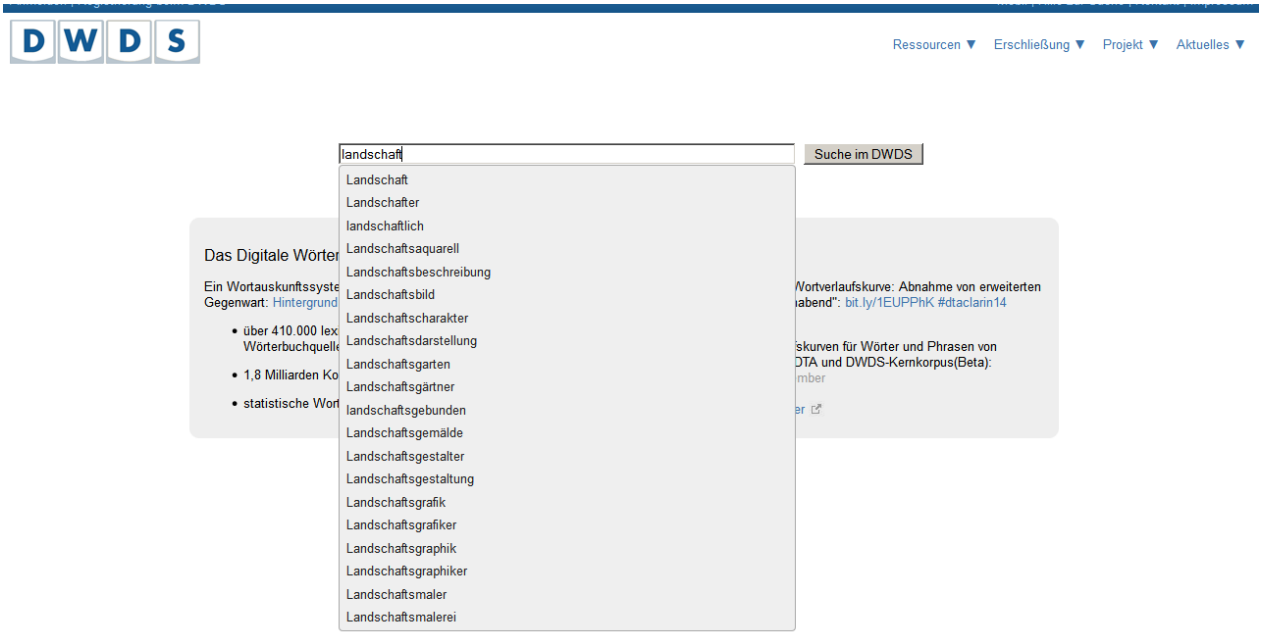
Tutorial zur Korpusrecherche mit DWDS

Über www.dwds.de können Sie sowohl auf das Kernkorpus des „Digitalen Wörterbuchs der Deutschen Sprache“ als auch auf das Deutsche Textarchiv (DTA) zugreifen, Ersteres mit Daten aus dem 20.Jh., Letzteres mit historischen Daten. Nicht nur wegen der sehr einfachen Bedienbarkeit (keine Anmeldung erforderlich, Export unproblematisch möglich) ist dies eine sehr wertvolle Ressource.

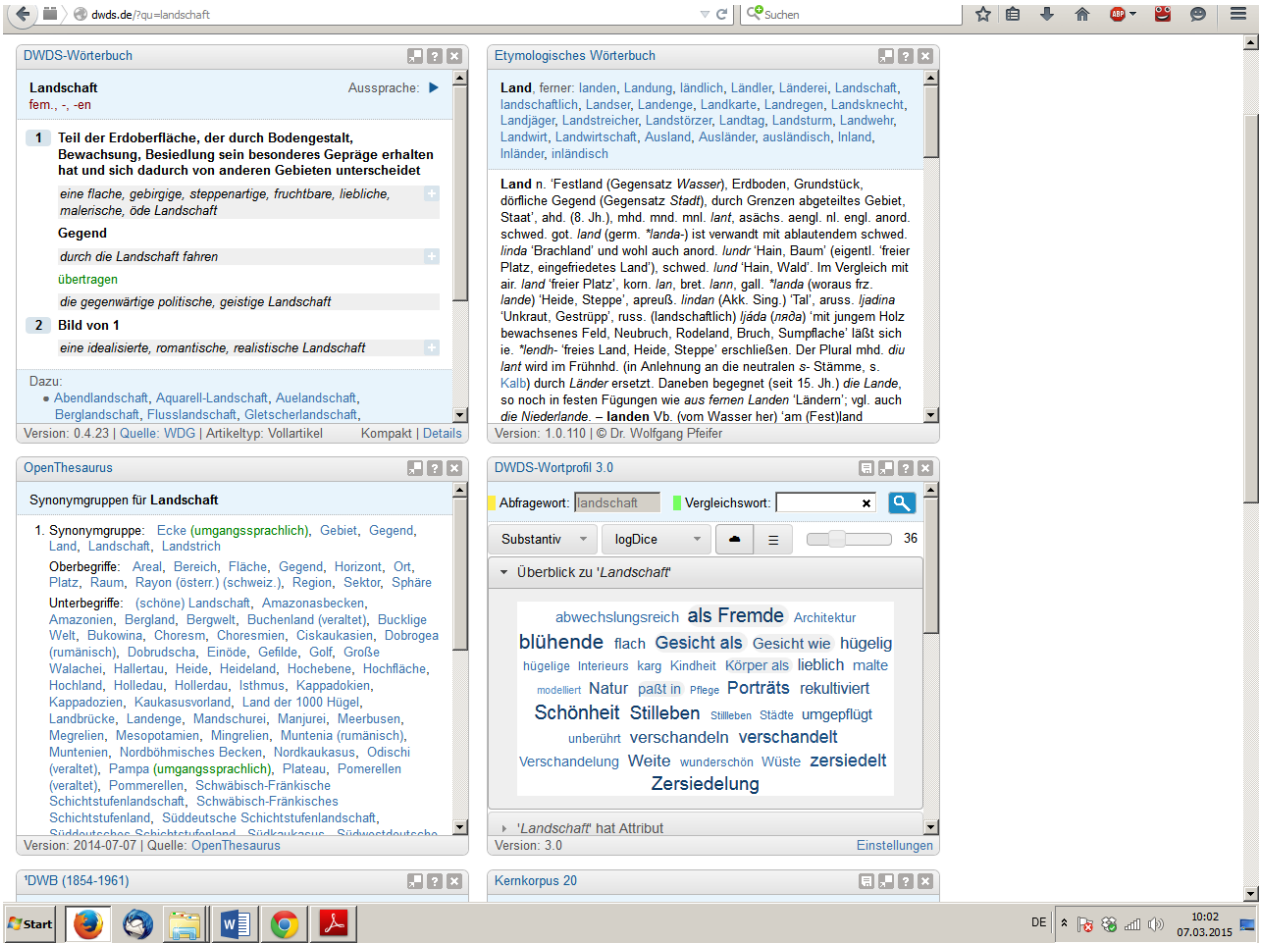
Update, November 2016: DWDS hat inzwischen ein neues Interface, das den Export leider auf gerade einmal 100 (!) Treffer beschränkt. Das alte Interface ist jedoch noch unter eins.dwds.de erreichbar. Dieses Tutorial bezieht sich ausschließlich auf das alte Interface.

1. Beispielanfrage: *-landschaft*

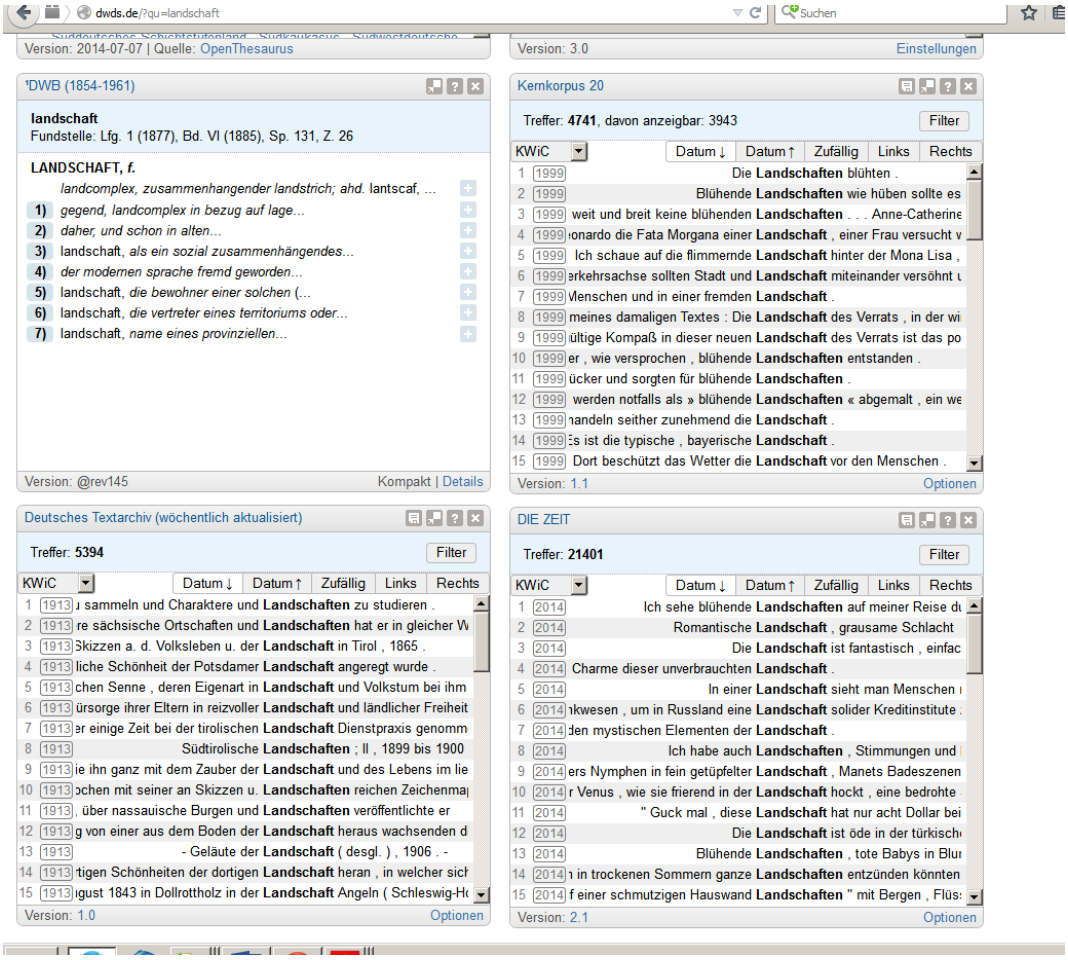
Angenommen, wir wollen den Gebrauch des Zweitglieds (Affixoids?) *-landschaft* untersuchen, z.B. *Hochschullandschaft*, *Korpuslandschaft*. Hierfür geben wir bei DWDS zunächst einfach „landschaft“ ein.



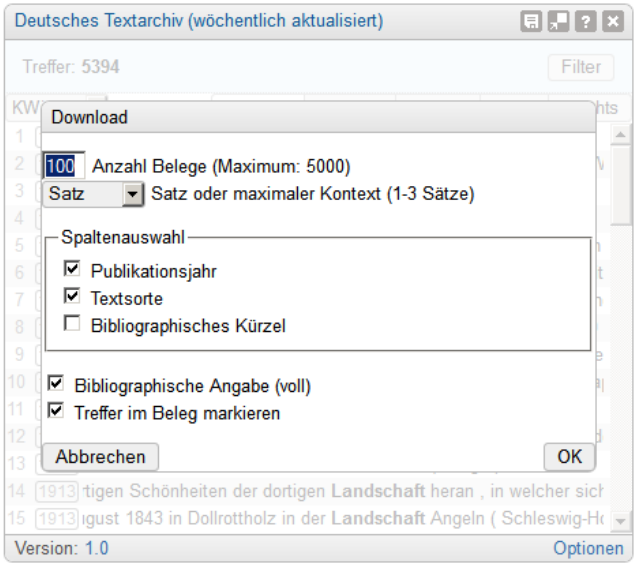
Wir gelangen auf eine Seite, in der Einträge aus mehreren Wörterbüchern und Resultate aus mehreren Korpora versammelt sind:



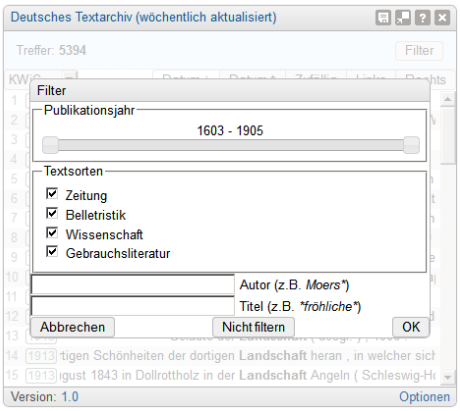
Oben rechts zum Beispiel sehen Sie den Eintrag aus dem Etymologischen Wörterbuch von Pfeifer, links Einträge aus dem DWDS-Wörterbuch und dem OpenThesaurus. Rechts unten das DWDS-Wortprofil, das auf Grundlage der Korpusdaten bestimmte Eigenschaften errechnet, die das gesuchte Wort / die gesuchte Wortgruppe häufig aufweist.
Zu den für uns interessanten Korpusdaten gelangen wir, wenn wir weiter nach unten scrollen.



Neben dem bereits erwähnten Kerkorpus und dem Deutschen Textarchiv gibt es auch ein Korpus aus Texten der Wochenzeitung DIE ZEIT. Für unsere Beispielanfrage arbeiten wir jedoch mit dem DTA, im obigen Screenshot links unten. Mit dem Button oben links am Fenster können Sie die Konkordanz herunterladen:



Wie Sie sehen, ist die Anzahl der zu exportierenden Belege auf 100 voreingestellt, was natürlich für eine quantitative Korpusanalyse geradezu lächerlich wenig ist (für eine qualitative Korpusanalyse hingegen sind 100 vollkommen ok). Auch sehen Sie, dass maximal 5000 Belege exportiert werden können – wir haben jedoch 5394 Treffer, d.h. wir müssen entweder 394 unter den Tisch fallen lassen oder die Daten in zwei Blöcke aufteilen. Der Vollständigkeit halber entscheiden wir uns für Letzteres. Mit einem Klick auf **Abbrechen** gelangen wir zum Korpusfenster zurück und klicken oben rechts auf die Schaltfläche **Filter.**



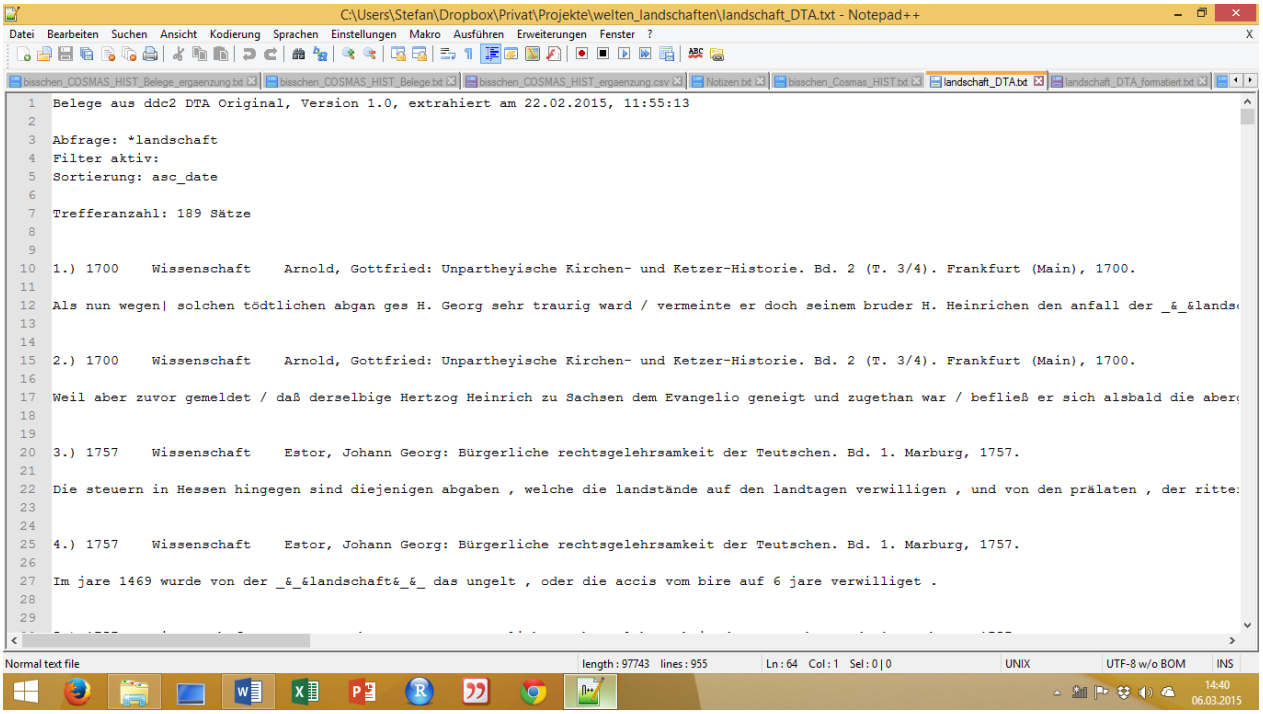
Nun gibt es mehrere Möglichkeiten, die Daten aufzuteilen, z.B. zuerst Zeitungs- und belletristische Texte zu exportieren, dann Wissenschaft und Gebrauchsliteratur; oder aber einen Filter nach Publikationsjahr zu setzen. Wir entscheiden uns hier für Letzteres und setzen einen Filter von 1603-1830. Jetzt sind noch 3006 Treffer, die sich problemlos exportieren lassen. Wir klicken wieder auf **Download**, stellen die **Anzahl der Belege** auf 5000 (oder 3006) ein und setzen bei „Spaltenauswahl“ Häkchen bei **Publikationsjahr** und **Textsorte**, sodass beide Metainformationen in der exportierten Konkordanz auftauchen. Nun können wir die Konkordanz als .txt-Datei herunterladen.

2. Übertragen des KWIC in ein Spreadsheet

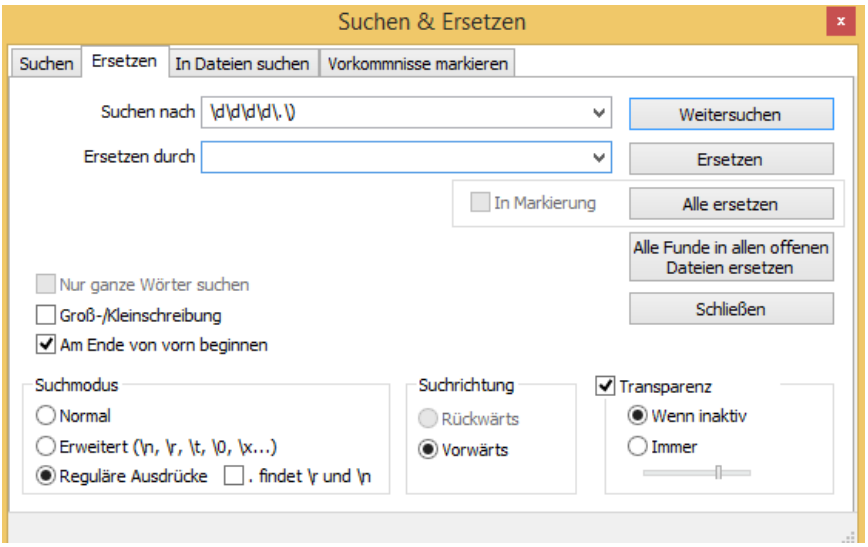
Wir öffnen die Datei mit **Notepad++**, einem kostenlosen Editor, der unter http://notepad-plus-plus.org/ heruntergeladen werden kann. (Notepad++ gibt es leider nur für Windows. Als Alternative für Mac-User ist z.B. **TextWrangler** empfehlenswert.)

Es ist recht schnell erkennbar, dass die Dateien, die DWDS exportiert, nicht primär für die Weiterverarbeitung in Tabellenkalkulationsprogrammen konzipiert sind. Mit einigen Tricks lassen sich aber auch diese Konkordanzen in Excel übertragen.

Schauen wir uns zunächst die Struktur der Exportdateien an. Zunächst findet sich ein Header, der die Suchanfrage, die Anzahl der Treffer usw. angibt. Diesen können wir für die weitere Verarbeitung getrost löschen.



Wenn wir den Header gelöscht haben, beginnt unser Dokument mit *1.)*. Die Nummerierung der Belege brauchen wir aber nicht unbedingt, daher entfernen wir sie. Im Feld Suchen > Ersetzen (oder Strg+H) aktivieren wir hierfür die Option „**Reguläre Ausdrücke**“.



Wir ersetzen nun alle Fälle, in denen eine vierstellige Ziffernfolge von einem Punkt und einer schließenden Klammer gefolgt wird, durch nichts (Feld „Ersetzen durch“ einfach leer lassen). Mit regulären Ausdrücken finden wir diese Fälle wie folgt:

\d\d\d\d\.\)

Das gleiche wiederholen wir für dreistellige Ziffernfolgen:

\d\d\d\.\)

...für zweistellige...

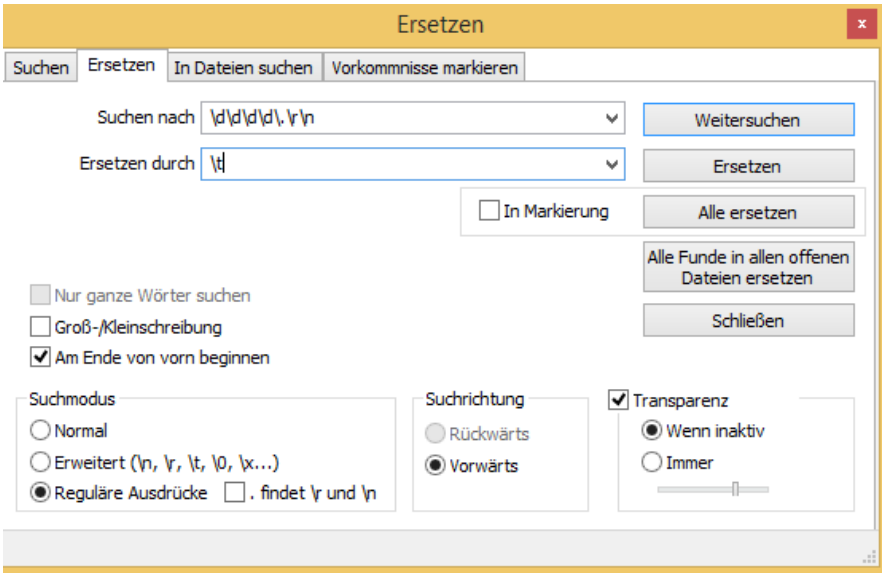
\d\d\.\)

...und schließlich für einstellige:

\d\.\)

Leere Zeilen löschen
Als nächstes wollen wir die leeren Zeilen loswerden. Dies geht in aktuellen Versionen von Notepad++ erfreulich einfach:
Bearbeiten > Zeilenoperationen > Leere Zeilen (auch mit Whitespace) löschen

Eine Zeile - ein Beleg
Nun ist aber immer noch jeder Beleg auf zwei Zeilen verteilt. Wie können wir das ändern? Erfreulicherweise haben die Belege eine klare Struktur: Die erste Zeile enthält immer die Quellenangabe, die in DTA mit einer Jahreszahl endet, beim DWDS-Kernkorpus mit einer Seitenangabe. Dann folgt der eigentliche Beleg.
Um beide in eine Zeile zu bekommen, getrennt durch einen Tab, verwenden wir reguläre Ausdrücke. Am einfachsten ist dies für DTA mit Jahreszahlen:



Wir ersetzen alle \d\d\d\d\.\r\n durch \t. Wieder steht \d für eine Ziffer, \d\d\d\d also für eine vierstellige Ziffernfolge (Jahreszahl). \r\n findet den Zeilenumbruch.
Dadurch werden zwar die Jahreszahlen am Ende der Quellenangabe getilgt, aber wir haben ja bereits eine eigene Spalte mit den Jahreszahlen.
Da einige Jahreszahlen in DTA unklar sind, stehen sie in eckigen Klammern, daher müssen wir außerdem noch alle \[\d\d\d\d\]\.\r\n durch \t ersetzen. Anschließend unbedingt manuell nachkorrigieren, da möglicherweise nicht alle Fälle korrekt erkannt wurden.
Da bei DWDS zumeist Seitenzahlen am Ende stehen, ersetzen wir hier \d\r\n durch \t. Auch hier gilt: Unbedingt manuell nachkorrigieren!

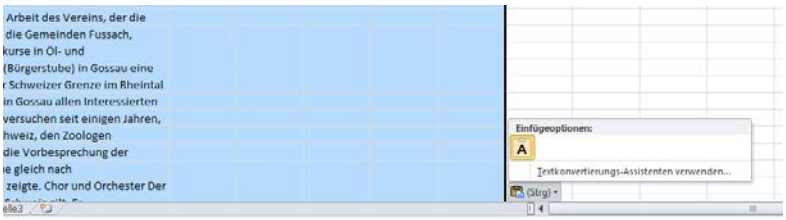
Keyword in eigene Spalte
Sehr positiv an den DTA/DWDS-Exportdateien ist, dass das Keyword durch eine Zeichenfolge abgegrenzt ist, das in authentischem Text wohl nie auftritt, sodass beim Ersetzen nicht mit Fehltreffern zu rechnen ist. Sie sind mit _&_& bzw. &_&_ abgetrennt. Diese Zeichenfolgen müssen einfach durch Tabs (\t) ersetzt werden.

Außerdem...
Um Fehler beim Copy&Pasten zu vermeiden, empfiehlt es sich außerdem, alle Anführungszeichen, Bindestriche und ggf. auch Semikola durch nichts zu ersetzen, sofern sie für Ihre Daten nicht relevant sind (und meistens sind sie das nicht).

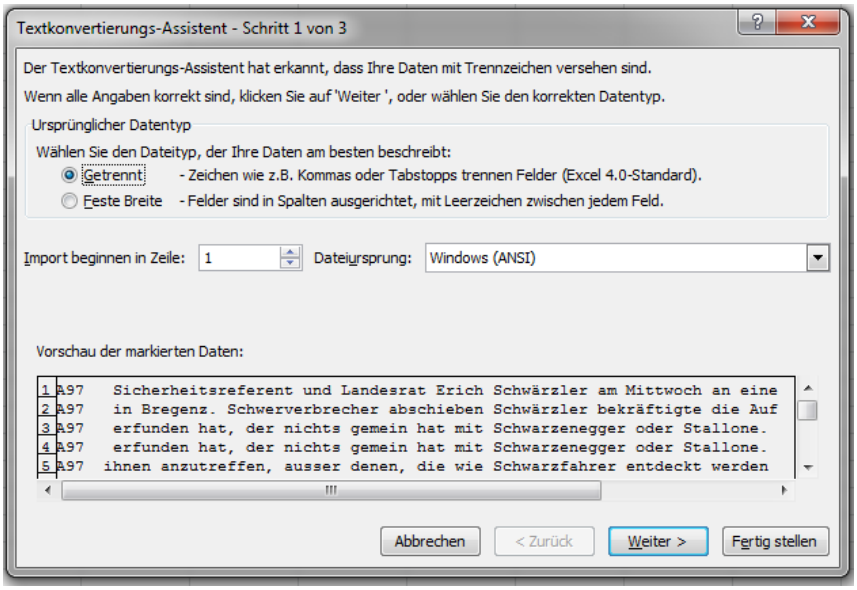
3. Zu guter Letzt
Nun können Sie die Daten ganz einfach in Excel copy&pasten. Dabei sollten Sie, um sicherzugehen, dass Excel die Dateien richtig interpretiert, die Schritte befolgen, die im COSMAS II-Tutorial angegeben wurden – gerade dann, wenn Sie die Anführungszeichen nicht gelöscht haben. Hier noch einmal der Text aus dem COSMAS II-Tutorial:

Hinweis: Der folgende Text wurde 1:1 aus dem COSMAS-Tutorial übernommen und bezieht sich deshalb auf andere Beispiele (hier: Erstglied/Präfixoid *schwarz-* in z.B. *schwarzfahren*)

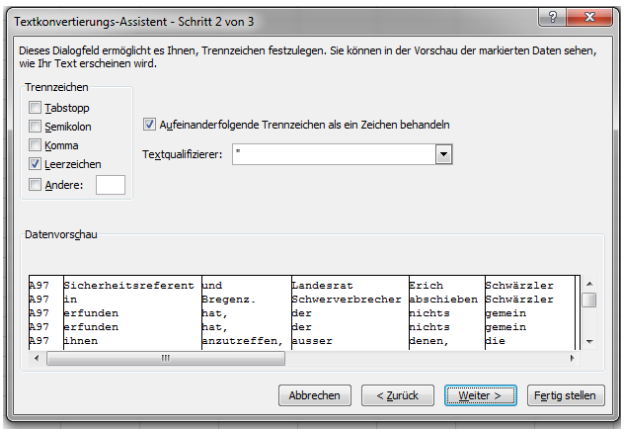
Wir klicken gleich nach dem Einfügen auf das kleine Kästchen neben der Markierung, in dem „{Strg}“ steht (dieses verschwindet wieder, sobald wir die Markierung aufheben, deshalb gleich draufklicken!).

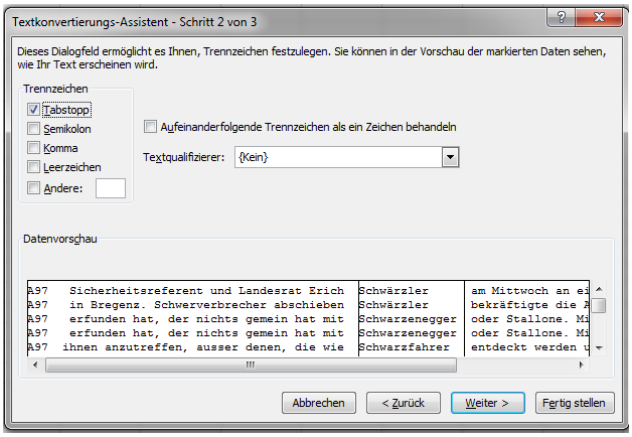


Hier wählen wir die Option „Textkonvertierungs-Assistenten verwenden“. Nun können wir Excel in drei Schritten mitteilen, wie es die Daten interpretieren soll:

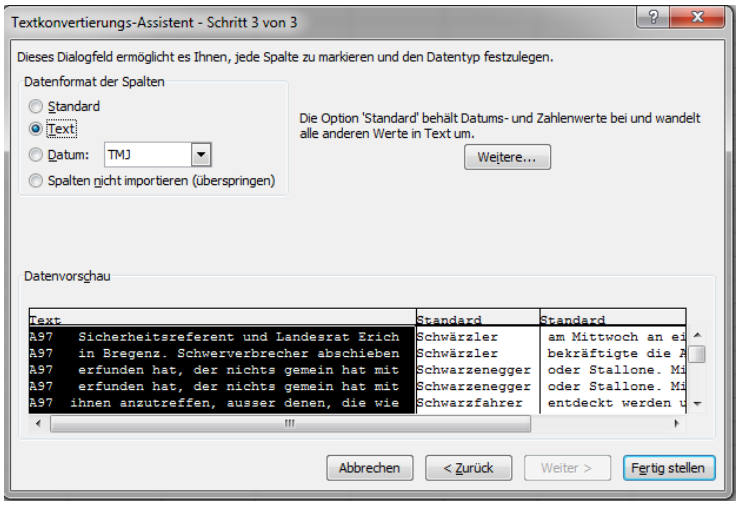


Im ersten Schritt ist die von uns gewollte Option „Getrennt“ bereits angewählt, wir können also einfach auf „Weiter“ klicken.





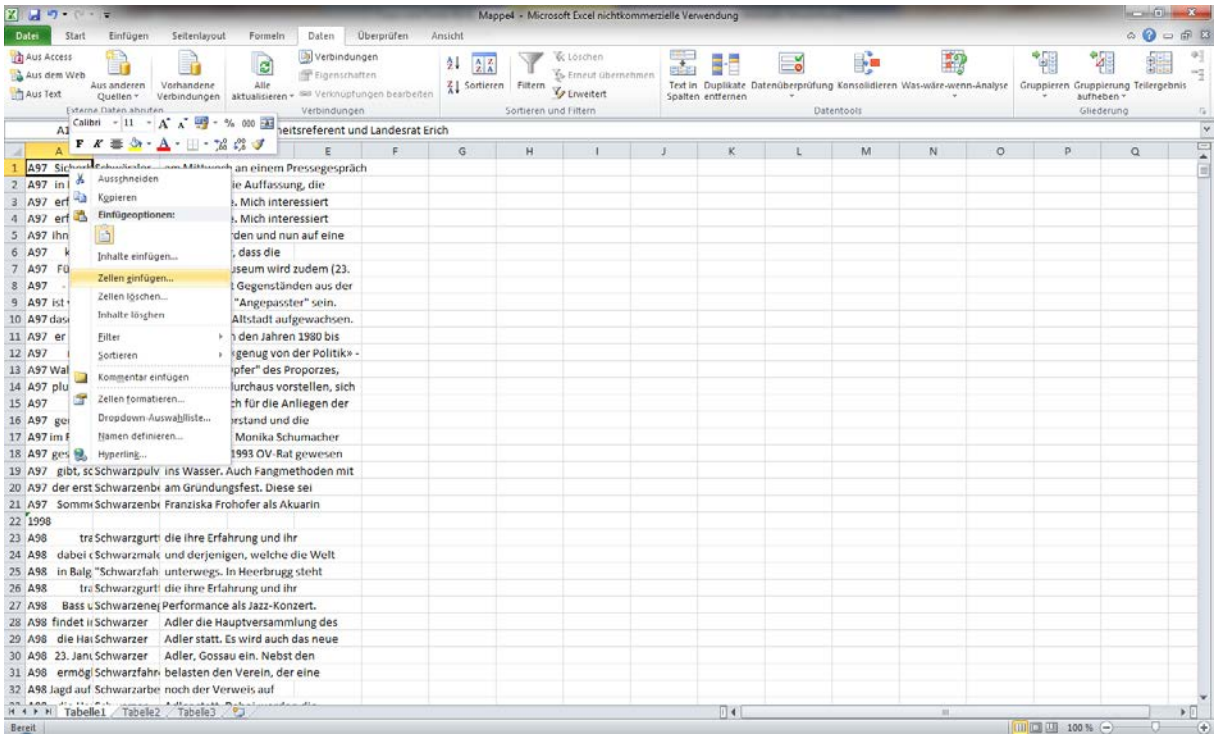
Im zweiten Schritt wählen wir **Tabstopp** als Trennzeichen und - ganz wichtig - deaktivieren die Erkennung von Anführungszeichen als Textqualifizierer, indem wir als **Textqualifizierer** *{kein}* auswählen.



Im dritten Schritt schließlich können wir Excel noch mitteilen, dass es sich bei unseren Daten um **Text** handelt (nicht etwa um Zahlen oder um Kalenderdaten). Abschließend klicken wir auf **Fertig stellen.**

Arbeit mit dem Excel-Dokument

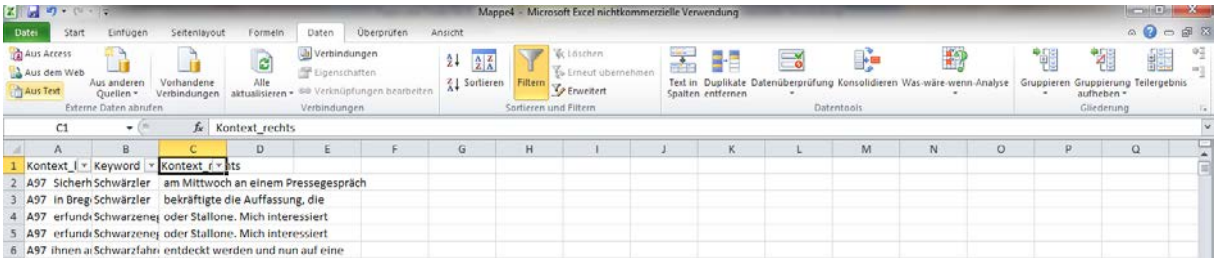
Geschafft: Wir haben unsere **Konkordanz**, also unsere Belegsammlung, erfolgreich in Excel eingefügt. Um die Daten im Excel-Dokument nach Keyword sortieren zu können, fügen wir zunächst eine Überschriftenzeile ein: Rechtsklick in der ersten Zeile > Zellen einfügen > **Ganze Zeile.**



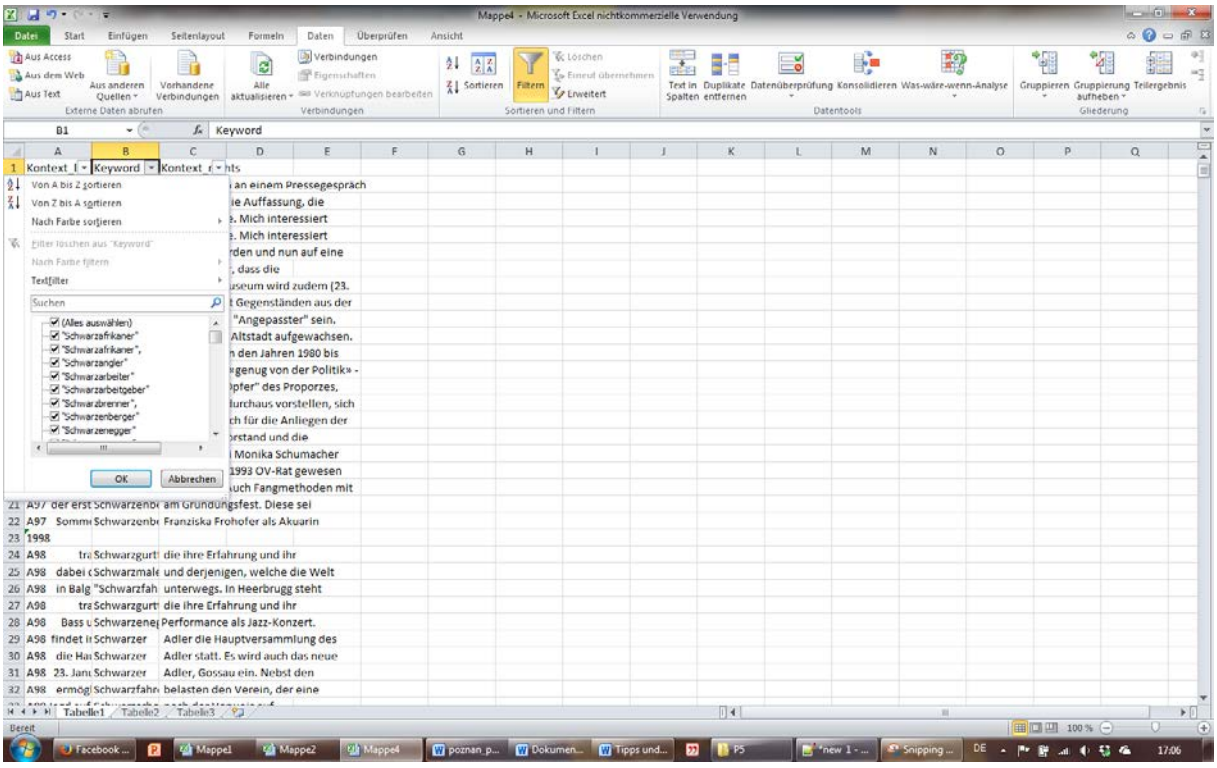
In die neu entstandene Zeile tragen wir nun die Überschriften für die einzelnen Spalten ein, z.B. Kontext_links, Keyword, Kontext_rechts.

	A	B	C	D	E
1	Kontext_link	Keyword	Kontext_rechts		
2	A97 Sicherh	Schwärzler	am Mittwoch an einem Pressegespräch		
3	A97 in Breg	Schwärzler	bekräftigte die Auffassung, die		
4	A97 erfunde	Schwarzeneg	oder Stallone. Mich interessiert		
5	A97 erfunde	Schwarzeneg	oder Stallone. Mich interessiert		
6	A97 ihnen ai	Schwarzfahr	entdeckt werden und nun auf eine		
7	A97 kündi	Schwärzler	will nicht nur, dass die		
8	A97 Fürster	Schwarzenbe	Gemeindemuseum wird zudem (23.		
9	A97 - für N	Schwarzer.	Amriswil: Mit Gegenständen aus der		
10	A97 ist vielle	Schwarzer	(45) will kein "Angepasster" sein.		
11	A97 dasein D	Schwarzer	ist in Arbons Altstadt aufgewachsen.		
12	A97 er sich	Schwarzer	war bereits in den Jahren 1980 bis		
13	A97 nach	Schwarzer	einstweilen «genug von der Politik» -		
14	A97 Wahlen	Schwarzer	wurde ein "Opfer" des Proporzes,		
15	A97 plus Sta	Schwarzer	könnte sich durchaus vorstellen, sich		
16	A97 me	Schwarzer	hauptberuflich für die Anliegen der		

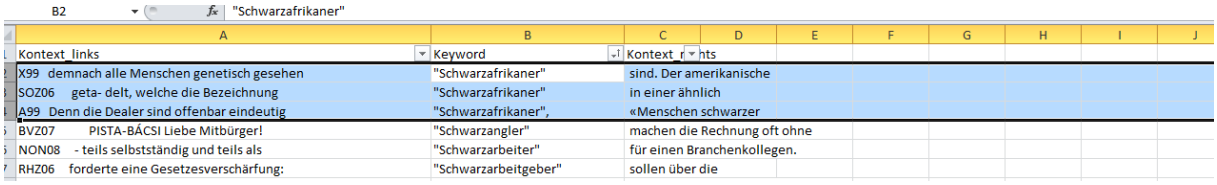
Wir lassen die erste Zeile markiert, gehen ganz oben auf den Reiter DATEN und dort auf **Filtern**.



Nun sehen Sie kleine Pfeilchen neben den Überschriften. Mit einem Klick auf das Pfeilchen neben **Keyword** können wir die Schlüsselwörter „Von A bis Z sortieren“:

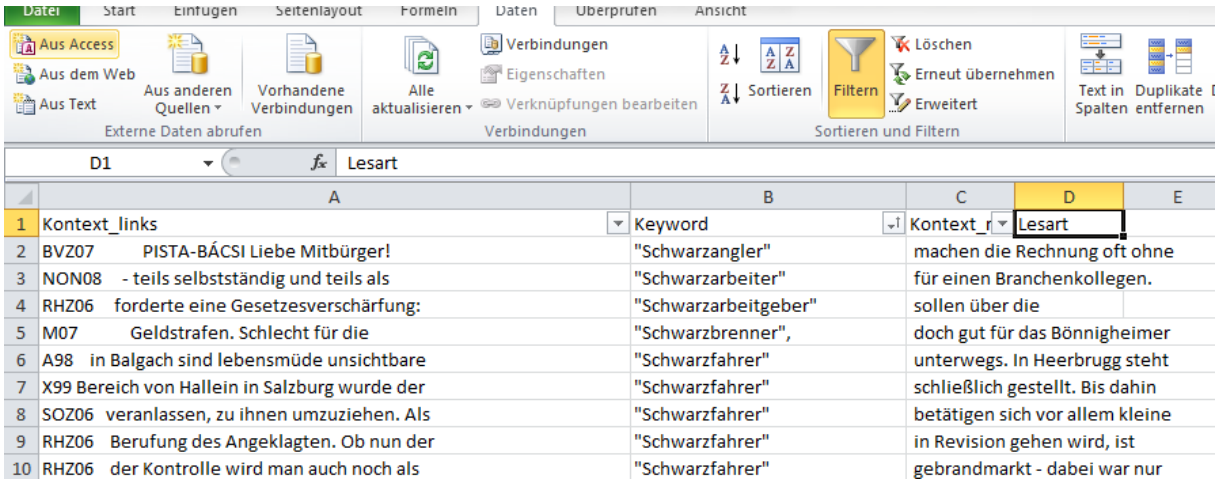


Damit haben wir unsere Liste de facto nach **Types** sortiert und können Fehltreffer bequem aussortieren. Tipp: Mit **Strg + Leertaste** markieren Sie **die ganze Zeile**, mit Strg + **- löschen** Sie den markierten Bereich. Die ersten drei Belege in der hier vorliegenden Konkordanz lauten *Schwarzafrikaner* (das ist in Ihrer Konkordanz wegen anderer Zufallsauswahl evtl. anders). Das ist zwar eine Personenbezeichnung, aber trotz ihres rassistischen Gehalts und der damit einhergehenden negativen Konnotation keine produktive Verwendung von *Schwarz-*, die uns interessieren würde. Deshalb markieren wir die ersten drei Zeilen (mit Umschalt + Pfeiltasten können Sie mehrere Zeilen markieren) und dehnen die Markierung mit Hilfe der oben genannten Tastenkombination **Strg + Leertaste** dann auf die **gesamten Spalten** aus.



Kontext_links	Keyword	Kontext_r…hts
X99 demnach alle Menschen genetisch gesehen	"Schwarzafrikaner"	sind. Der amerikanische
SOZ06 geta- delt, welche die Bezeichnung	"Schwarzafrikaner"	in einer ähnlich
A99 Denn die Dealer sind offenbar eindeutig	"Schwarzafrikaner",	«Menschen schwarzer
BVZ07 PISTA-BÁCSI Liebe Mitbürger!	"Schwarzangler"	machen die Rechnung oft ohne
NON08 - teils selbstständig und teils als	"Schwarzarbeiter"	für einen Branchenkollegen.
RHZ06 forderte eine Gesetzesverschärfung:	"Schwarzarbeitgeber"	sollen über die

Nun können wir uns mit Hilfe von **Strg** + - der Fehltreffer entledigen. Die nächsten Belege, *Schwarzangler*, *Schwarzarbeiter* und *Schwarzarbeitgeber* sind hingegen genau die Verwendungen, die wir suchen. Bald kommen jedoch mit mehreren Belegen für *Schwarzer* wieder Fehltreffer. Hier können wir das eben skizzierte Vorgehen wiederholen, ebenso für alle anderen Fehltreffer, derer es hier sehr viele gibt. Umso besser, dass wir sie mit nur drei Tastenkombinationen schnell und effizient beseitigen können! Was übrigbleibt, ist eine recht überschaubare Belegsammlung, die wir ggf. noch näher annotieren wollen. Zum Beispiel könnten wir in einer Spalte festhalten, wie das *Schwarz-* jeweils zu interpretieren ist: Bezieht es sich auf eine „Kostenumgehung“ wie z.B. *Schwarzfahrer* oder *Schwarzpinkler* oder auf eine pessimistische Einstellung wie *Schwarzmaler* oder *Schwarzseher* (Letzteres jedoch ambig durch wortspielhaften Gebrauch z.B. in GEZ-Werbespots)?
Wir fügen also eine Spalte „Lesart“ an:



	Kontext_links	Keyword	Kontext_r	Lesart
2	BVZ07 PISTA-BÁCSI Liebe Mitbürger!	"Schwarzangler"	machen die Rechnung oft ohne	
3	NON08 - teils selbstständig und teils als	"Schwarzarbeiter"	für einen Branchenkollegen.	
4	RHZ06 forderte eine Gesetzesverschärfung:	"Schwarzarbeitgeber"	sollen über die	
5	M07 Geldstrafen. Schlecht für die	"Schwarzbrenner",	doch gut für das Bönnigheimer	
6	A98 in Balgach sind lebensmüde unsichtbare	"Schwarzfahrer"	unterwegs. In Heerbrugg steht	
7	X99 Bereich von Hallein in Salzburg wurde der	"Schwarzfahrer"	schließlich gestellt. Bis dahin	
8	SOZ06 veranlassen, zu ihnen umzuziehen. Als	"Schwarzfahrer"	betätigen sich vor allem kleine	
9	RHZ06 Berufung des Angeklagten. Ob nun der	"Schwarzfahrer"	in Revision gehen wird, ist	
10	RHZ06 der Kontrolle wird man auch noch als	"Schwarzfahrer"	gebrandmarkt - dabei war nur	

Diese ist zunächst nicht gefiltert, aber indem wir den Filter oben deaktivieren und dann wieder aktivieren, stellen wir sicher, dass die gesamte Zeile gefiltert wird und wir auch die Lesart, wenn wir die entsprechende Spalte annotiert haben, alphabetisch sortiert und eben auch gefiltert werden kann.
Wichtig bei der Annotation, wenn Sie quantitativ an die Daten herangehen wollen: Arbeiten Sie mit wenigen, klar definierten **Variablenausprägungen**. Was heißt das? Ganz einfach: In unserem Fall ordnen Sie jeden Beleg z.B. der Ausprägung ILLEGAL oder der Ausprägung PESSIMISTISCH zu – auch wenn Sie bei einigen Belegen vielleicht Bauchschmerzen haben, sie einer der beiden Kategorien zuzuordnen.

	Kontext_links	Keyword	Kontext_rechts	Lesart
2	BVZ07 PISTA-BÁCSI Liebe Mitbürger!	"Schwarzangler"	machen die Rechnung oft ohne	ohne Genehmigung angeln
3	NON08 - teils selbstständig und teils als	"Schwarzarbeiter"	für einen Branchenkollegen.	arbeiten ohne Steuern zu zahlen
4	RHZ06 forderte eine Gesetzesverschärfung:	"Schwarzarbeitgeber"	sollen über die	Mensch der jemandem Arbeit gibt der keine Steuern zahlt
5	M07 Geldstrafen. Schlecht für die	"Schwarzbrenner",	doch gut für das Bönnigheimer	ich hab auch schon viele Raubkopien gebrannt ist aber verjährt
6	A98 in Balgach sind lebensmüde unsichtbare	"Schwarzfahrer"	unterwegs. In Heerbrugg steht	hier wohl Wortspiel
7	X99 Bereich von Hallein in Salzburg wurde der	"Schwarzfahrer"	schließlich gestellt. Bis dahin	warum die Anführungszeichen
8	SOZ06 veranlassen, zu ihnen umzuziehen. Als	"Schwarzfahrer"	betätigen sich vor allem kleine	ich hab keine Lust mehr
9	RHZ06 Berufung des Angeklagten. Ob nun der	"Schwarzfahrer"	in Revision gehen wird, ist	ich hab Hunger
10	RHZ06 der Kontrolle wird man auch noch als	"Schwarzfahrer"	gebrandmarkt - dabei war nur	
11	RHZ06 der Kontrolle wird man auch noch als	"Schwarzfahrer"	gebrandmarkt - dabei war nur	
12	HMP06 Ticket hatte, brachten Polizisten den	"Schwarzfahrer"	in einen Reptilienzoo.	
13	RHZ06 fanden neben dem achtbeinigen	"Schwarzfahrer"	einen Zettel an den "lieben	

SO NICHT!

Am Ende haben Sie dann – hoffentlich – ein übersichtliches Dokument mit Ihren Belegen und Annotationen, das Sie quantitativ auswerten können. Dafür gibt es in Excel ebenfalls einige hilfreiche Funktionen (mit denen ich mich aber weniger gut auskenne, da ich eher mit dem Statistikprogramm R arbeite). Ich hoffe jedoch, dass Ihnen dieses Tutorial den Einstieg in die Korpuslinguistik etwas einfacher gemacht hat. Für Verbesserungsvorschläge zu diesem Dokument bin ich natürlich jederzeit dankbar!