

Tutorial zur Sortierung von COSMAS II-Belegen nach Jahr

Wenn Sie chronologisch sortierte Belege aus COSMAS II exportieren¹, haben die Resultate folgendes Format (vgl. COSMAS II-Tutorial) - hier am Beispiel des Keywords *ein wenig*:

```
Jahr unbekannt:
HK4 Frauenklubs sind. Dürfen wir uns nicht ein wenig rühmen, wir alle, die wir
HK4 ihm nicht verborgen, er mokierte sich ein wenig darüber. Nun hätte es wohl
HK4 indem er seine Schwerhörigkeit ein wenig übertrieb. Denn ließ die Arbeit eine
HK4 seinem Antlitz. Sie atmete die fremde, ein wenig nach Moder schmeckende Luft, die um
HK4 väterlich herab und stellten sie auf die ein wenig bebenden Kniee. Hat sich mein
HK4 und sie wies auf den Alten. Der stand ein wenig blöd und zitternd mit vorgeschobnem
HK4 schönen Strauß dieser Blüten öffnete es ein wenig sein weißliches Lid. Mag sein, daß es
HK4 all das stimmte ihn nachdenklich und fast ein wenig so wie ein Gespräch mit einem guten
HK4 er fort, da steht man auf, und er reckt ein wenig die Arme. Er lächelt wieder mit
HK4 ins Gesicht. Er denkt, ich sei ein wenig sonderbar, gleichviel man soll nicht
HK4 an, wenn am Samstag der Zug in die Berge ein wenig vor Büroschluß abgeht. Er hat mit
HK4 er schon älter über die Wünsche hinaus, ein wenig griesgrämig und mit Sorgen um sein
HK4 Würde und Sicherheit des Auftretens sich ein wenig ins Lächerliche begeben. Im
HK4 einer Bürste blonden, an den Schläfen ein wenig gelockten Haares. Er wandte sich
HK4 im Insel-Verlag) zu vergleichen, ist der ein wenig schmerzliche Einblick in eine Seele,
HK4 hat darin, nach meinem Gefühl, zuweilen ein wenig zu viel getan. Wo aber, wie hier,
1658
GMC es ist der Gebrauch in Indien/ daß wer ein wenig vornehm ist/ jhm lasset so wol auff
GMC auch zwene Kauffleute mit Lantzen ein wenig versehret. Wir reyseten mit der
1660
GMC mancherley Gedancken für. Wann er gleich ein wenig ruhet/ so ists doch nichts/ dann er
1662
HK4 damit sie schneller schnellert / das Glas ein wenig trübt / auf daß die Herzen Perl /
HK4 will / wie jener Thor es macht / der mit ein wenig Erd die Sonne wolt verstreichen. Ihr
HK4 und es finden in der That. 8. Laß mir nur ein wenig leuchten / deine Weißheit / mich
HK4 erniedern selbst / sie pflaget zuerheben / ein wenig kränkt / auf daß sie wider neu
1664
GMC sind viel Sitz-Bäncke/ und dieselbigen ein wenig voneinander unterschieden/ und mit
.....
```

Mit Hilfe dieses Tutorials (und eines R-Skripts, siehe unten) können Sie diese Daten in eine Tabelle überführen, in der die Jahreszahlen - die im KWIC-Export als Überschriften fungieren - in einer eigenen Korpuspalte vermerkt sind.

Für die Korpusrecherche mit Excel und den Export der Daten verweise ich auf mein **COSMAS II**-Tutorial. Das aktuelle Tutorial setzt an dem Punkt an, an dem Sie die Daten aus Cosmas exportiert haben und **bevor** Sie sie in Excel bearbeiten (das passiert im anderen Tutorial in Punkt 7).

Sie benötigen:

- Notepad++ (siehe Cosmas-Tutorial)
- R
- (optional, aber sehr zu empfehlen) R Studio

¹ Denken Sie daran, dass Sie zweischrittig vorgehen müssen, um eine nach Jahr sortierte KWIC-Tabelle aus COSMAS zu exportieren: 1.) Im Reiter „KWIC“ die Daten chronologisch sortieren, 2.) im Reiter „Export“ aus dem Dropdown-Menü „KWIC, chronologisch sortiert“ auswählen. Schritt 2) ist nur möglich, wenn Sie vorher Schritt 1) durchgeführt haben.

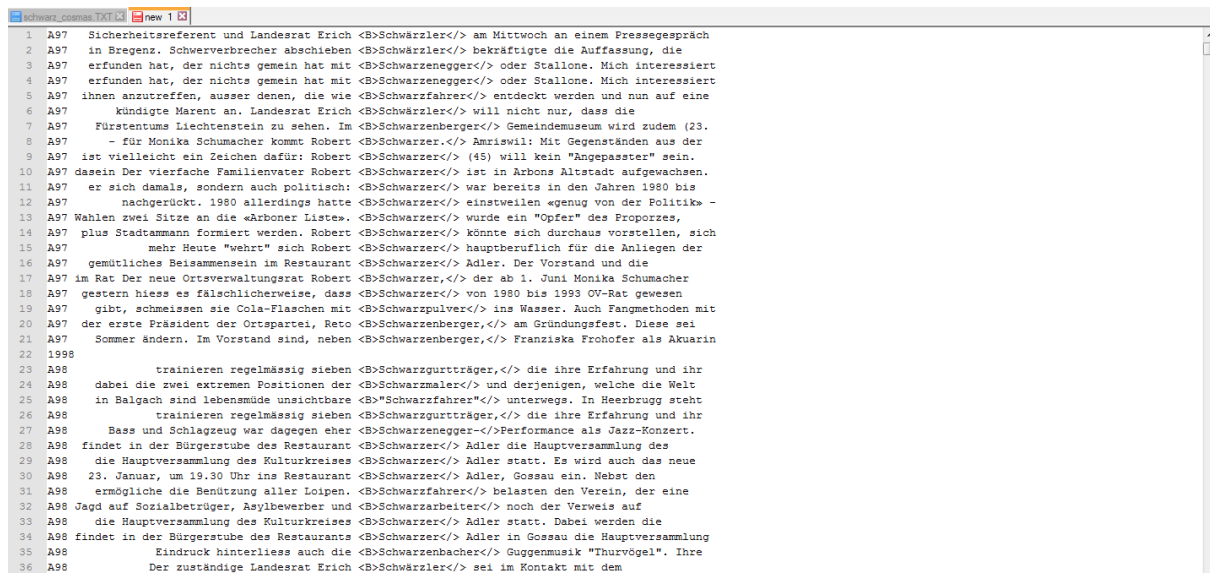
1. Vorbereiten der Exportdatei

Die Exportdatei ist, wie ebenfalls im COSMAS-Tutorial vermerkt, dreigeteilt:

- 1.) Header mit der Suchanfrage und Ergebnisübersicht
- 2.) KWIC
- 3.) Volltext

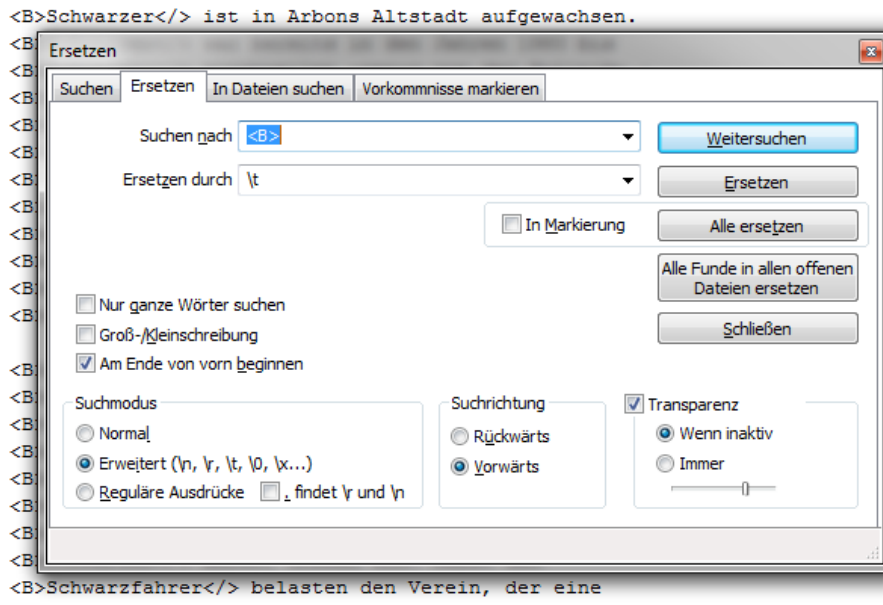
Wichtig: Um die folgenden Schritte in diesem Tutorial korrekt durchführen zu können, ist es wichtig, dass Sie zunächst **nur** (!) den KWIC-Teil in ein neues Dokument copy&pasten und dieses Dokument abspeichern. Der KWIC-Teil beginnt mit dem ersten Jahr (im Screenshot oben: „Jahr unbekannt“) und endet mit dem letzten Beleg. Den Beginn des Volltext-Teils erkennen Sie an der Überschrift „Belege (chronologisch sortiert)“.

Mit dem neuen Textdokument, das nur die KWIC-Daten enthält (nennen wir es einfach **corpus_kwic.txt** - natürlich können Sie auch jeden anderen Dateinamen verwenden), arbeiten wir nun weiter (in Notepad++ - auch hier gilt: vgl. COSMAS-Tutorial).

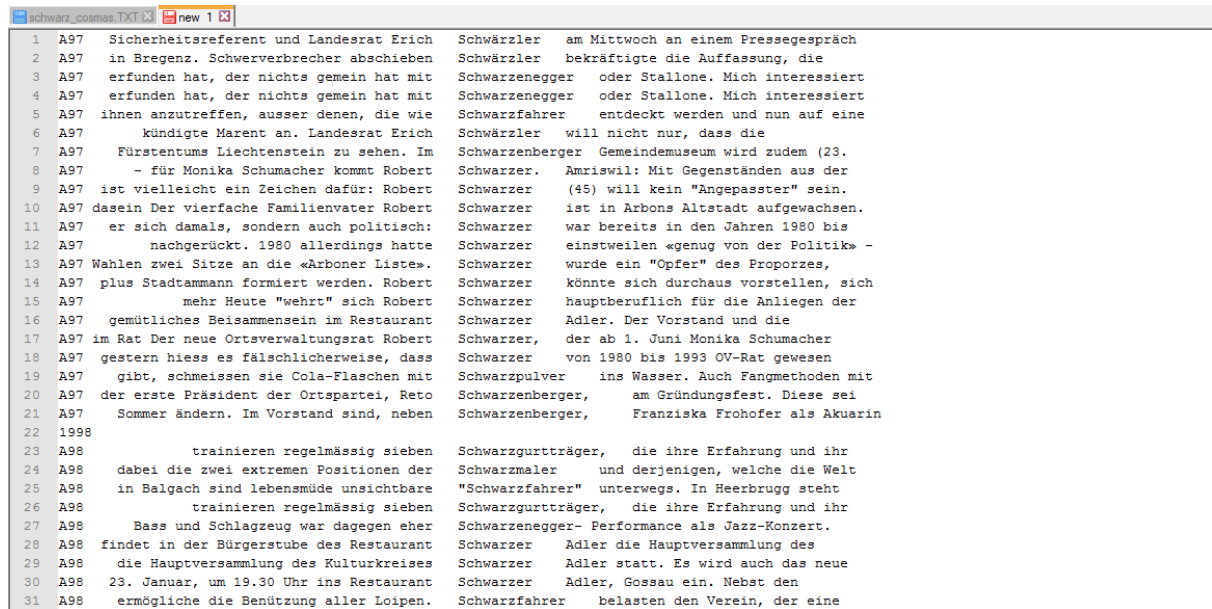


```
schwarz_cosmas.TXT [new 1]
1 A97 Sicherheitsreferent und Landesrat Erich <B>Schwarzler</> am Mittwoch an einem Pressegespräch
2 A97 in Bregeuz. Schwerverbrecher abschieben <B>Schwarzler</> bekräftigte die Auffassung, die
3 A97 erfunden hat, der nichts gemein hat mit <B>Schwarzenegger</> oder Stallone. Mich interessiert
4 A97 erfunden hat, der nichts gemein hat mit <B>Schwarzenegger</> oder Stallone. Mich interessiert
5 A97 ihnen anzutreffen, ausser denen, die wie <B>Schwarzfahrer</> entdeckt werden und nun auf eine
6 A97 kündigte Marent an. Landesrat Erich <B>Schwarzler</> will nicht nur, dass die
7 A97 Fürstentums Liechtenstein zu sehen. Im <B>Schwarzenberger</> Gemeindegemüse wird zudem (23.
8 A97 - für Monika Schumacher kommt Robert <B>Schwarzer.</> Amriswil: Mit Gegenständen aus der
9 A97 ist vielleicht ein Zeichen dafür: Robert <B>Schwarzer</> (45) will kein "Angepasster" sein.
10 A97 dasein Der vierfache Familienvater Robert <B>Schwarzer</> ist in Arbons Altstadt aufgewachsen.
11 A97 er sich damals, sondern auch politisch: <B>Schwarzer</> war bereits in den Jahren 1980 bis
12 A97 nachgerückt. 1980 allerdings hatte <B>Schwarzer</> einstweilen «genug von der Politik» -
13 A97 Wahlen zwei Sitze an die «Arboner Liste». <B>Schwarzer</> wurde ein "Opfer" des Proporz,
14 A97 plus Stadtmann formiert werden. Robert <B>Schwarzer</> könnte sich durchaus vorstellen, sich
15 A97 mehr heute "wehrt" sich Robert <B>Schwarzer</> hauptberuflich für die Anliegen der
16 A97 gemütliches Beisammenssein im Restaurant <B>Schwarzer</> Adler. Der Vorstand und die
17 A97 im Rat Der neue Ortsverwaltungsrat Robert <B>Schwarzer,</> der ab 1. Juni Monika Schumacher
18 A97 gestern hiess es fälschlicherweise, dass <B>Schwarzer</> von 1980 bis 1993 OF-Rat gewesen
19 A97 gibt, schmeissen sie Cola-Flaschen mit <B>Schwarzpulver</> ins Wasser. Auch Fangmethoden mit
20 A97 der erste Präsident der Ortspartei, Reto <B>Schwarzenberger,</> am Gründungsfest. Diese sei
21 A97 Sommer ändern. Im Vorstand sind, neben <B>Schwarzenberger,</> Franziska Frohofer als Aduarin
22 1998
23 A98 trainieren regelmässig sieben <B>Schwarzgutträger,</> die ihre Erfahrung und ihr
24 A98 dabei die zwei extremen Positionen der <B>Schwarzmaier</> und derjenigen, welche die Welt
25 A98 in Balgach sind lebensmüde unsichtbare <B>Schwarzfahrer</> unterwegs. In Heerbrugg steht
26 A98 trainieren regelmässig sieben <B>Schwarzgutträger,</> die ihre Erfahrung und ihr
27 A98 Bass und Schlagzeug war dagegen eher <B>Schwarzenegger</>Performance als Jazz-Konzert.
28 A98 findet in der Bürgerstube des Restaurant <B>Schwarzer</> Adler die Hauptversammlung des
29 A98 die Hauptversammlung des Kulturkreises <B>Schwarzer</> Adler statt. Es wird auch das neue
30 A98 23. Januar, um 19.30 Uhr ins Restaurant <B>Schwarzer</> Adler, Gossau ein. Nebst den
31 A98 ermögliche die Benützung aller Loipen. <B>Schwarzfahrer</> belasten den Verein, der eine
32 A98 Jagd auf Sozialbetrüger, Asylbewerber und <B>Schwarzarbeiter</> noch der Verweis auf
33 A98 die Hauptversammlung des Kulturkreises <B>Schwarzer</> Adler statt. Dabei werden die
34 A98 findet in der Bürgerstube des Restaurants <B>Schwarzer</> Adler in Gossau die Hauptversammlung
35 A98 Eindruck hinterliess auch die <B>Schwarzenbacher</> Guggenmusik "Thurvögel". Ihre
36 A98 Der zuständige Landesrat Erich <B>Schwarzler</> sei im Kontakt mit dem
```

Wie Sie sehen, ist das **Keyword** (hier ein anderes Beispiel als zuvor, da der Screenshot direkt aus dem COSMAS II-Tutorial übernommen wurde) in **** und **</>** eingeschlossen. Um die Daten in ein Tabellenkalkulationsprogramm (z.B. Excel) zu exportieren, **ersetzen** wir diese beiden Marker durch **Tabs**. Über *Suchen>Ersetzen* oder einfach Strg+H gelangen wir in das Ersetzungsfenster.



Stellen Sie sicher, dass als Suchmodus „erweitert“ ausgewählt ist: Nur so können wir \t als Code für Tabs verwenden. Wir ersetzen nun alle durch \t und wiederholen dies für </> (jeweils mit „Alle ersetzen“, dritter Button von oben). Die Datei sieht nun so aus:



Weitere, optionale Schritte:

- Falls für Ihre Recherche unwichtig, ersetzen Sie alle **Anführungszeichen** und alle **Bindestriche** im Dokument durch nichts (Ersetzungsfeld einfach leer lassen). Das macht später den Import in Excel etwas einfacher.
- Für Fortgeschrittene, die im weiteren Verlauf der Recherche mit Excel-generierten .csv-Dateien arbeiten wollen (z.B. in R): Ersetzen Sie auch alle **Semikola** (;) durch nichts, weil die sonst später fälschlicherweise als Trennzeichen erkannt werden könnten.

2. Bearbeiten der KWIC-Datei mit R

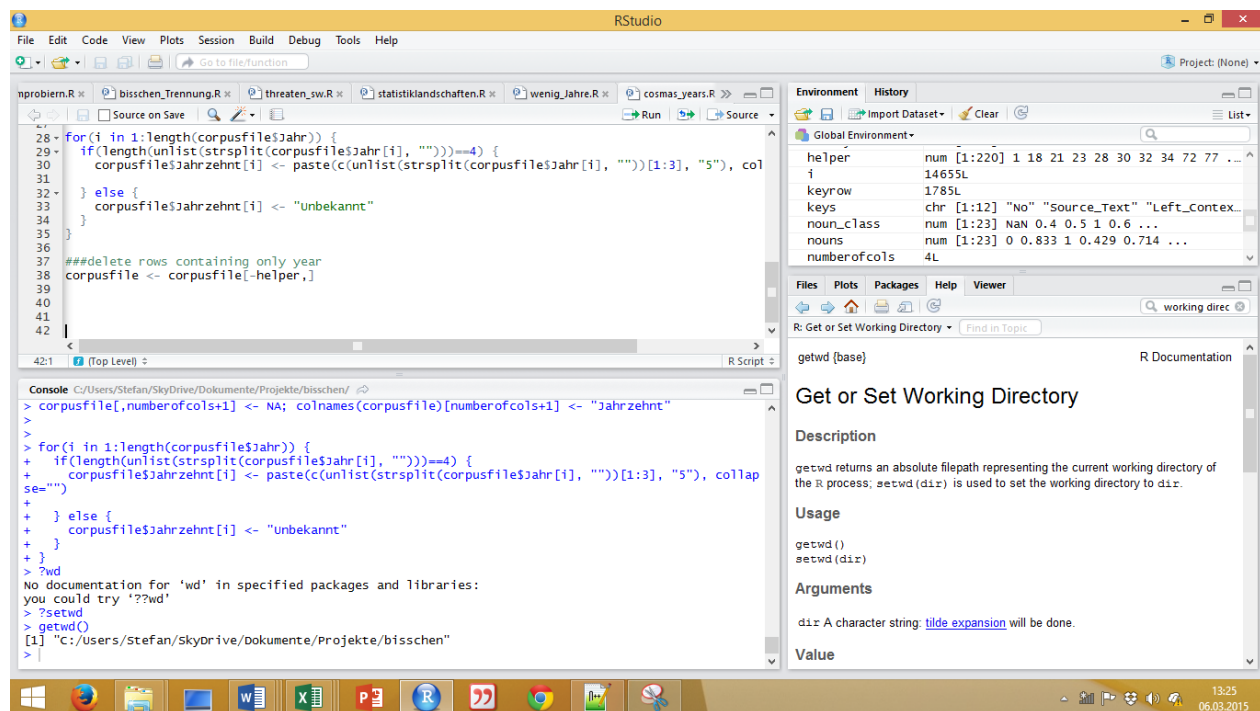
Für das Skript zur Bearbeitung der KWIC-Datei habe ich die Open-Source-Programmiersprache R benutzt (hauptsächlich, weil es bis jetzt die einzige ist, die ich halbwegs beherrsche). An dieser Stelle kann ich keine Einführung in R geben und auch den Code im Skript nicht umfassend erklären. Auch kann ich nicht garantieren, dass das Skript mit jeder Input-Datei funktioniert. Ich gehe aber davon aus, dass das Skript in den meisten Fällen brauchbare Ergebnisse erzeugt.

2.1 R und R Studio installieren

Die Installation von R und R Studio ist sehr einfach und selbsterklärend. R erhalten Sie unter www.r-project.org, R Studio unter www.r-studio.com. R und R Studio gibt es für die meisten gebräuchlichen Betriebssysteme.

2.2 Skript benutzen

Öffnen Sie R Studio. Der Bildschirm ist viergeteilt:



Das Fenster oben links entspricht praktisch einem Texteditor, hier können Sie Skripts schreiben. Das Fenster unten links ist die **Konsole**, in der die Befehle, die Sie eingeben, ausgeführt werden. Die beiden anderen Fenster sind für unsere Zwecke uninteressant.

Wichtig ist: Wenn Sie im Editorfenster oben links einen **Befehl** eingeben, können Sie ihn mit **Strg+Eingabe** an die Konsole übertragen, wo er ausgeführt wird. Probieren Sie:

```
sqrt(25)
```

Wenn Sie Strg+Eingabe drücken, erhalten Sie die Wurzel aus 25, also 5.

Wenn am Anfang einer Zeile `##` steht, interpretiert R die Zeile nicht als Befehl. Notizen in einem Skript beginnen daher immer mit `##`. Probieren Sie:

```
##sqrt(25)
```

Der Text wird zwar an die Konsole übertragen, aber es wird kein Befehl ausgeführt.

Nun können Sie ansatzweise den Aufbau des **Skripts** nachvollziehen, das Sie herunterladen können (`cosmas_years.R`) und das Sie auch unten im Anhang als Text finden.

Bevor Sie das Skript öffnen, finden Sie heraus, was das aktuelle **Arbeitsverzeichnis** von R ist.

Tippen Sie:

```
getwd()
```

Wenn Sie mögen, können Sie das Arbeitsverzeichnis mit `setwd(HIER DATEIPFAD EINSETZEN)` ändern.

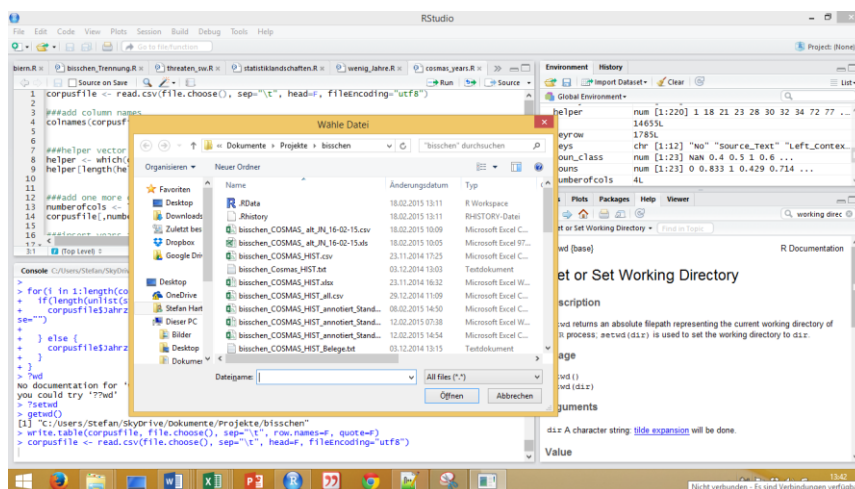
Copy&pasten Sie die Textdatei mit den KWIC-Daten (für die ich oben den Namen `corpus_kwic.txt` vorgeschlagen habe) sowie das Skript `cosmas_years.R` in das Arbeitsverzeichnis von R.

Öffnen Sie nun das Skript `cosmas_years.R`.

Das Skript macht fast alles alleine und braucht nur zwei Angaben von Ihnen: In der allerersten und in der allerletzten Zeile müssen Sie jeweils eine Datei auswählen. Zeile 1:

```
corpusfile <- read.csv(file.choose(), sep="\t", head=F, fileEncoding="utf8")
```

Wenn Sie diese Zeile ausführen (mit `Strg+Eingabe` oder über den Button „Run“), öffnet sich ein Fenster (das manchmal nicht von selbst aufpoppt - in diesem Fall müssen Sie auf das neu erschienene, meist blinkende Element in Ihrer Taskleiste klicken), das Sie auffordert, eine Datei auszuwählen.



Hier wählen Sie nun `corpus_kwic.txt` (oder wie auch immer Sie die Textdatei mit den KWIC-Daten genannt haben).

Den Rest macht das Skript quasi von alleine - in den nächsten Abschnitten dieses Tutorials können Sie einen kurzen Blick hinter die Kulissen werfen, aber Sie können auch einfach nach diesem

Abschnitt aufhören zu lesen und sich über die Ergebnisse freuen, ohne genau zu wissen, wie sie zustande gekommen sind ;-)

Am einfachsten ist es, wenn Sie einfach den gesamten Rest des Skripts markieren und mit Strg+Eingabe oder dem Button „Run“ ausführen. In der allerletzten Zeile müssen Sie dann noch einmal ein Dokument auswählen:

```
write.table(corpusfile, file.choose(), sep="\t", row.names=F, quote=F)
```

Hier wählen Sie das Dokument, in das die Tabelle gespeichert werden soll, das R aus Ihren Daten gezaubert hat.

Vorsicht:

Sie können in diesem letzten Schritt auch schon existierende Dateien wählen.

Diese Dateien werden überschrieben!

R fragt nicht nach, bevor es die existierenden Dateien ersetzt!!

Geben Sie also am besten einen ganz neuen Dateinamen ein. Vergessen Sie außerdem die Dateiendung nicht (z.B. .txt für eine Textdatei, alternativ auch .csv für ein Spreadsheet, das Sie theoretisch direkt in Excel öffnen können - aber sicherer ist es, die Datei zunächst in Notepad++ zu öffnen und dann in Excel zu copy&pasten).

Nun können Sie wieder am Ende des COSMAS II-Tutorials ansetzen, wo erklärt wird, wie Sie eine tab-separierte Tabelle aus Notepad in Excel copy&pasten.

3. Hinter den Kulissen: Was macht das Skript?

Wenn R das KWIC-Dokument einliest, interpretiert das Programm es als Tabelle, wobei Tabs die Spalten voneinander abgrenzen – hier wieder das Beispiel aus dem allerersten Screenshot, der Pfeil steht für einen Tab:

```
HK4   Frauenklubs sind. Dürfen wir uns nicht   →   ein wenig   →   rühmen, für alle, die
```

Spalte 1 enthält also den linken Kontext, Spalte 2 das Keyword, nach dem wir bei der Korpusrecherche gesucht haben (hier: „ein wenig“), Spalte 3 den Kontext, der dem Keyword folgt. Das Skript gibt diesen drei Spalten zunächst sprechende Namen (`colnames`), nämlich „Kontext_links“, „Keyword“, „Kontext_rechts“.

Wie wir oben gesehen haben, fungieren die Jahreszahlen im ursprünglichen KWIC quasi als Überschriften. Das bedeutet, dass in den Zeilen, in denen eine Jahreszahl steht, **nur** diese Jahreszahl steht. Kein Keyword, kein rechter Kontext: Spalte 2 und 3 sind also leer. Das Skript macht sich diese Tatsache zunutze und sucht die Zeilen, in denen Spalte 2 und 3 leer sind.² Es kopiert die Jahreszahl in eine neu erstellte vierte Spalte, und zwar in jeder Zeile einmal, bis die nächste

² In einer früheren Version habe ich stattdessen nach denjenigen Zeilen gesucht, in denen in der Spalte „Kontext_links“ entweder vier Ziffern oder „Jahr unbekannt“ standen. Dies scheitert jedoch daran, dass aus Gründen, die ich noch näher untersuchen will, einige der Ziffern in den Jahreszahlen bei Verwendung nicht als Ziffern erkannt werden.

Jahreszahl in der ersten Spalte steht. Das wird so lange wiederholt, bis alle Jahreszahlen abgearbeitet sind.

Weil das Arbeiten mit Jahren manchmal zu feinkörnig ist, macht das Skript zusätzlich noch eine fünfte Spalte auf, in dem die Jahreszahlen auf Jahrzehnte heruntergebrochen werden. Dafür wird einfach die letzte Ziffer der Jahreszahl durch eine „5“ ersetzt, d.h. sowohl 1861 als auch 1869 gehören zum Jahrzehnt 1860-1869, das kurz als „1865“ angegeben wird.

Anhang: R-Skript

```
corpusfile <- read.csv(file.choose(), sep="\t", head=F, fileEncoding="utf8")

###add column names
colnames(corpusfile)[1:3] <- c("Kontext_links", "Keyword", "Kontext_rechts")

###helper vector: find years
helper <- which(corpusfile$Keyword==" " & corpusfile$Kontext_rechts==" ")
helper[length(helper)+1] <- length(corpusfile$Kontext_links)+1

###add one more column
numberofcols <- length(corpusfile)
corpusfile[,numberofcols+1] <- NA; colnames(corpusfile)[numberofcols+1] <-
"Jahr"

###insert years in new column
for(i in 1:(length(helper)-1)) {
  corpusfile[(helper[i]+1):(helper[i+1]-1),]$Jahr <- gsub("[:punct:]", "",
as.character(corpusfile[helper[i],1]))
}

###add decades
numberofcols <- length(corpusfile)
corpusfile[,numberofcols+1] <- NA; colnames(corpusfile)[numberofcols+1] <-
"Jahrzehnt"

for(i in 1:length(corpusfile$Jahr)) {
  if(length(unlist(strsplit(corpusfile$Jahr[i], "")))==4) {
    corpusfile$Jahrzehnt[i] <- paste(c(unlist(strsplit(corpusfile$Jahr[i],
"")))[1:3], "5"), collapse="")
  } else {
    corpusfile$Jahrzehnt[i] <- "Unbekannt"
  }
}

###delete rows containing only year
corpusfile <- corpusfile[-helper,]
write.table(corpusfile, file.choose(), sep="\t", row.names=F, quote=F)
```