Das R-Paket „concordances“

Die Exportdateien von Korpusabfragesystemen kommen oft nicht in einem Format, das es erlauben würde, sie direkt in ein Tabellenkalkulationsprogramm wie Excel einzulesen. Lange habe ich deshalb jede Konkordanz manuell bearbeitet. Da das recht zeitaufwendig ist, habe ich ein Paket für R geschrieben, das diesen Prozess automatisiert. Das Paket *concordances* ist zwar noch *work in progress*, aber zumindest für die Exportdateien des WaCkY-Korpus und von COSMAS II scheint es zu funktionieren. Bugs sind dennoch nicht ausgeschlossen und sogar wahrscheinlich - wenn etwas nicht funktioniert, teilen Sie es mir gerne mit!

Installation des Pakets
Da *concordances* (noch) nicht über das „Comprehensive R Architecture Network“ (CRAN) verfügbar ist, kann man es nicht, wie andere Pakete, einfach mit dem Befehl `install.packages()` installieren. Die einfachste Variante, das Paket zu installieren, ist, zunächst das Paket *devtools* zu installieren (dieses wiederum kann man mit install.packages() installieren) und anschließend die Funktion `install_github()` zu nutzen.

```
install.packages("devtools")
devtools::install_github("hartmast/concordances")
```

Laden des Pakets
Wenn ein Paket installiert ist, gibt es zwei Möglichkeiten, die darin enthaltenen Funktionen zu benutzen. Eine besteht darin, auf einzelne Funktionen zuzugreifen, indem man zuerst den Namen des Pakets eingibt, dann zwei Doppelpunkte und dann den Namen der Funktion, wie wir es gerade eben mit *devtools::* gemacht haben. Gerade wenn man Funktionen aus einem Paket in einem Skript öfter benutzt, ist es jedoch sinnvoller, das Paket zu laden. Dann kann man direkt auf die Funktionen zugreifen und muss nicht jedesmal den Namen des Pakets mit eingeben. Wenn Sie *concordances* erfolgreich installiert haben, können Sie es wie folgt laden:

```
library("concordances")
```

Benutzen des Pakets
Das Paket beinhaltet derzeit folgende Funktionen, die Sie nutzen können, sobald Sie es geladen haben:

- getCOSMAS: Hiermit lassen sich Konkordanzen, die aus COSMAS II exportiert wurden, als R-Dataframes einlesen. Unterstützt werden unsortierte KWIC-Konkordanzen und solche im Jahr- oder Jahrzehnformat.
- getWACKY: Diese Funktion liest XML-Exportdateien aus der NoSketchEngine-Installation des WACKY-Korpus ein. (Die Funktion sollte auch mit XML-Exportdateien aus anderen (No)SketchEngine-Korpora kompatibel sein; sie wird mittelfristig in der nächsten Funktion aufgehen.)
- getNSE: Diese Funktion liest .txt-Dateien ein, die mit NoSketchEngine erstellt wurden. Sie ist auf die Exportdateien des COW-Korpus abgestimmt, sollte aber auch mit anderen Konkordanzen kompatibel sein, die über NoSketchEngine (oder auch SketchEngine) erstellt wurden.

- getCWB: Diese Funktion liest Exportdateien der Corpus Workbench (CQP) ein. Hinweis: Wenn Sie mit CQP*web* arbeiten, erübrigt sich diese Funktion, denn dort können Sie Konkordanzen direkt in Spreadsheet-kompatiblem Format exportieren.

Erstellung von Spreadsheets mit *concordances*

In den meisten Fällen wollen wir die Daten, die wir haben, mit zusätzlichen Annotationen anreichern und benötigen die Konkordanz dafür in einem Format, das sich in einem Tabellenkalkulationsprogramm öffnen und bearbeiten lässt. Konkordanzen aus COSMAS II, CWB oder NoSketchEngine kann man sehr einfach als .csv-Spreadsheet speichern, das man dann z.B. in Excel oder Calc öffnen kann, indem man zunächst mit einer der oben genannten *get*-Funktionen die Exportdatei einliest und sie anschließend gleich wieder exportiert. Zum Export verwenden wir die R-Funktion *write.table*:

```
txt <- getCOSMAS("dateiname.txt")

write.table(txt, file = "konkordanz.csv", sep = "\t", row.names = F, quote = F)
```

In der ersten Zeile wird einfach die Datei eingelesen, wobei sie in dem R-Objekt mit dem Namen *txt* gespeichert wird. Daraufhin erfolgt der Export, wobei im obigen Beispiel für *write.table* folgende Argumente angegeben sind:

- file = "konkordanz.csv" gibt an, dass das Spreadsheet, das wir exportieren, unter dem Dateinamen *konkordanz.csv* gespeichert werden soll. Man kann hier auch einen ganzen Pfad angeben, z.B. C:/Stefan/Dokumente/konkordanz.csv. Ansonsten wird die Datei standardmäßig im aktuellen Arbeitsverzeichnis gespeichert, das sich mit getwd() erfragen und mit setwd() verändern lässt. Wichtig: Existiert die Datei bereits, wird sie ohne Rückfrage überschrieben!
- sep = "\t" gibt an, dass Tabstopps als Trennzeichen verwendet werden sollen. Auch andere Trennzeichen sind möglich, aber nicht alle sind sinnvoll: Satzzeichen wie Kommas oder Semikola sollte man nicht verwenden, weil sie sicherlich in den Korpusbelegen vorkommen.
- row.names = F gibt an, dass die Zeilennamen nicht mit exportiert werden sollen. In den Dataframes, die die *get*-Funktionen in *concordances* generieren, haben die Zeilen gar keine Namen, sondern sind einfach durchnummeriert, deshalb können wir auf diese Information verzichten. Außerdem rutschen die Spaltennamen nach links, wenn wir die row.names mit exportieren, was ein weiterer Grund ist, es nicht zu tun.
- quote = F schließlich sagt, dass keine Anführungszeichen als Texttrenner benutzt werden sollen. Es ist zwar in vielen Fällen sinnvoll, solche Trennzeichen zu verwenden, aber da die Konkordanz vermutlich in vielen Fällen Anführungszeichen enthält, neige ich dazu, sicherheitshalber darauf zu verzichten, damit es beim Import in Tabellenkalkulationsprogramme oder beim späteren Reimport in R nicht zu Problemen kommt. Generell ist *quote = T* vor allem dann sinnvoll, wenn das Zeichen, das als Trennzeichen verwendet wird – hier also der Tabstopp, s.o. – auch in der Konkordanz selbst vorkommt. Das sollte in aller Regel aber nicht der Fall sein.

Zu den einzelnen Funktionen

Nähere Informationen dazu, wie die einzelnen Funktionen verwendet werden können, erhalten Sie (in englischer Sprache) über die Hilfe-Datei der jeweiligen Funktion, die Sie aufrufen können, indem Sie ein Fragezeichen eingeben, gefolgt vom Namen der Funktion, z.B. `?getCOSMAS`.
Die Funktionen funktionieren derzeit noch sehr unterschiedlich, in Kürze werde ich aber hoffentlich Zeit finden, sie zu vereinheitlichen. Hier knapp die einzelnen Argumente der Funktionen:

getCOSMAS

- filename: Hier wird der Pfad zur Datei eingegeben, die eingelesen werden soll (also der Datei, die Sie aus COSMAS exportiert haben). Liegt die Datei im Arbeitsverzeichnis (das sich mit *getwd()* erfragen und mit *setwd()* ändern lässt), genügt der Dateiname. Achtung: Es muss der genaue Dateiname angegeben werden, Groß- und Kleinschreibung sind dabei relevant.
- merge: Wenn das Keyword aus mehr als einem Wort besteht (also genau genommen eine Key*phrase* ist), können Sie hier entscheiden, ob die Keywords in einer Spalte zusammengeführt werden oder in unterschiedliche Spalten aufgeteilt werden sollen. Ein Beispiel: Sie haben nach *je X-er desto Y-er* gesucht. Der Treffer besteht also nicht nur aus einem Wort, sondern aus vier (oder mehr, wenn Sie zusätzlich Abstandsoperatoren benutzt haben, um Fälle wie *je länger und mehr desto besser* einzubeziehen). Wenn `merge=FALSE` ist (der Default), dann werden die Keywords nicht zusammengeführt, sondern in unterschiedliche Spalten aufgeteilt, so wie auch in der Original-Exportdatei innerhalb eines Belegs jeder Treffer einzeln mit <B> ... </> hervorgehoben wird (siehe Screenshot in Fig. 1). Je nachdem, was genau man mit den Daten vorhat, kann das durchaus sinnvoll sein. Fig. 2 zeigt, wie die Konkordanz dann aussieht: Da es teilweise bis zu fünf Treffer pro Beleg gibt, gibt es insgesamt fünf Keyword-Spalten. In vielen Belegen gibt es aber weniger Treffer, die verbleibenden Keyword-Spalten sind dann NA (= Not Available). Mit `merge=TRUE` hingegen werden alle Wörter in einer Spalte zusammengeführt, wie in Fig. 3 dargestellt.
- years: COSMAS II bietet die Option, die Konkordanz im KWIC-Format unsortiert oder aber in einer Jahres- oder Jahrzehntansicht zu exportieren. Wurde eine der beiden letzteren Optionen gewählt, sollte man `years = TRUE` setzen, um sicherzustellen, dass die Jahreszahlen nicht in einer eigenen Zeile stehenbleiben (wie in der Original-Konkordanz, siehe Fig. 1), sondern eine eigene Spalte bekommen (Fig. 3). *getCOSMAS* fügt außerdem, wenn `years=TRUE` ist, automatisch eine Spalte „Decade“ hinzu, das die Jahreszahl auf die Mitte des jeweiligen Jahrzehnts rundet. Bei der Jahrzehntansicht ist diese Spalte natürlich überflüssig; arbeitet man mit dieser, kann man die Decade-Spalte anschließend z.B. im Tabellenkalkulationsprogramm einfach löschen.

```
1795
GOE     Wahlspruch wird doch nächstens sein:' <B>je</> weiter weg, <B>desto</> besser'. Sie werden mich,
GOE       er ist eine öffentliche Person, und <B>je</> ausgebildeter seine Bewegungen, <B>je</> sonorer
GOE         diese Rätsel beunruhigten mich um <B>desto</> mehr, <B>je</> mehr ich wünschte, zugleich
GOE         das schöne Kind und das Frühstück <B>je</> eher <B>je</> lieber entfernt zu sehen, und die
1808
GOE              je dringender das Bedürfnis, <B>je</> höher das Ahndungsvermögen, <B>je</> froher das
GOE      je beschränkter der Erkenntniskreis, <B>je</> dringender das Bedürfnis, <B>je</> höher das
1809
GOE dem Bade aus, und warf den ganzen Plunder <B>desto</> entschiedener von mir, <B>je</> mehr ich zu
GOE  das Dämonische so gern wirkt und uns nur <B>desto</> schlimmer mitspielt, <B>je</> mehr wir Ahndung
GOE         je tiefer der Winter sich senkte, <B>je</> wilderes Wetter, <B>je</> unzugänglicher die Wege,
1813
GOE  es geht mit der Kunst wie mit dem Leben: <B>je</> weiter man hineinkommt, <B>je</> breiter wird sie.
GOE   Geschrei so vieler Menschen, die nur um <B>desto</> heftiger brüllen, <B>je</> weniger sie ein
GOE    mais je ne suis rien et je sais ce que <B>je</> ne voudrois etre.<B> je</> souhaite, Monsieur, que
1819
GRI hatte. Der Wirt schrie zum Erbarmen, aber <B>je</> lauter er schrie, <B>desto</> kräftiger schlug der
GRI      my dat Äten schöön! Gif my mehr!" Un <B>je</> mehr he eet, <B>je</> mehr wull he hebben, un säd
GRI  Mann, "meine Natur ist ganz anderer Art, <B>je</> heißer es ist, <B>desto</> mehr frier ich, und der
1827
```

Fig. 1: Screenshot einer Original-Exportdatei, die man aus COSMAS II erhält. Innerhalb des einzelnen Belegs ist jeder Treffer, der dem Gesuchten entspricht, durch <B> ... </> hervorgehoben.

	A	B	C	D	E	F	G	H
1	Source	Left	Key1	Key2	Key3	Key4	Key5	Right
2	GOE	sagen, daß die organischen Naturen nur	desto	vollkommner werden,	je	NA	NA	weniger die
3	GOE	Wahlspruch wird doch nächstens sein:'	je	weiter weg,	desto	NA	NA	besser'. Sie werden mich,
4	GOE	er ist eine öffentliche Person, und	je	ausgebildeter seine Bewegungen,	je	NA	NA	sonorer
5	GOE	diese Rätsel beunruhigten mich um	desto	mehr,	je	NA	NA	mehr ich wünschte, zugleich
6	GOE	das schöne Kind und das Frühstück	je	eher	je	NA	NA	lieber entfernt zu sehen, und die
7	GOE	je dringender das Bedürfnis,	je	höher das Ahndungsvermögen,	je	NA	NA	froher das
8	GOE	je beschränkter der Erkenntniskreis,	je	dringender das Bedürfnis,	je	NA	NA	höher das
9	GOE	dem Bade aus, und warf den ganzen Plunder	desto	entschiedener von mir,	je	NA	NA	mehr ich zu
10	GOE	das Dämonische so gern wirkt und uns nur	desto	schlimmer mitspielt,	je	NA	NA	mehr wir Ahndung
11	GOE	je tiefer der Winter sich senkte,	je	wilderes Wetter,	je	NA	NA	unzugänglicher die Wege,
12	GOE	es geht mit der Kunst wie mit dem Leben:	je	weiter man hineinkommt,	je	NA	NA	breiter wird sie.
13	GOE	Geschrei so vieler Menschen, die nur um	desto	heftiger brüllen,	je	NA	NA	weniger sie ein
14	GOE	mais je ne suis rien et je sais ce que	je	ne voudrois etre.	je	NA	NA	souhaite, Monsieur, que
15	GRI	hatte. Der Wirt schrie zum Erbarmen, aber	je	lauter er schrie,	desto	NA	NA	kräftiger schlug der
16	GRI	my dat Äten schöön! Gif my mehr!" Un	je	mehr he eet,	je	NA	NA	mehr wull he hebben, un säd
17	GRI	Mann, "meine Natur ist ganz anderer Art,	je	heißer es ist,	desto	NA	NA	mehr frier ich, und der
18	GOE	bedarf weniger der Neuheit, ja vielmehr	je	älter sie ist,	je	NA	NA	gewohnter man sie ist,

Fig. 2: Mit getCOSMAS() erstelltes Spreadsheet mit der Option merge = FALSE: Eine aus mehreren Wörtern bestehende Keyphrase ist in einzelne Spalten aufgeteilt.

	A	B	C	D	E	F
1	Source	Left	Key	Right	Year	Decade
2	GOE	sagen, daß die organischen Naturen nur	desto vollkommner werden, je	weniger die	1790	1795
3	GOE	Wahlspruch wird doch nächstens sein:'	je weiter weg, desto	besser'. Sie werden mich,	1795	1795
4	GOE	er ist eine öffentliche Person, und	je ausgebildeter seine Bewegungen, je	sonorer	1795	1795
5	GOE	diese Rätsel beunruhigten mich um	desto mehr, je	mehr ich wünschte, zugleich	1795	1795
6	GOE	das schöne Kind und das Frühstück	je eher je	lieber entfernt zu sehen, und die	1795	1795
7	GOE	je dringender das Bedürfnis,	je höher das Ahndungsvermögen, je	froher das	1808	1805
8	GOE	je beschränkter der Erkenntniskreis,	je dringender das Bedürfnis, je	höher das	1808	1805
9	GOE	dem Bade aus, und warf den ganzen Plunder	desto entschiedener von mir, je	mehr ich zu	1809	1805
10	GOE	das Dämonische so gern wirkt und uns nur	desto schlimmer mitspielt, je	mehr wir Ahndung	1809	1805
11	GOE	je tiefer der Winter sich senkte,	je wilderes Wetter, je	unzugänglicher die Wege,	1809	1805
12	GOE	es geht mit der Kunst wie mit dem Leben:	je weiter man hineinkommt, je	breiter wird sie.	1813	1815
13	GOE	Geschrei so vieler Menschen, die nur um	desto heftiger brüllen, je	weniger sie ein	1813	1815
14	GOE	mais je ne suis rien et je sais ce que	je ne voudrois etre. je	souhaite, Monsieur, que	1813	1815
15	GRI	hatte. Der Wirt schrie zum Erbarmen, aber	je lauter er schrie, desto	kräftiger schlug der	1819	1815
16	GRI	my dat Äten schöön! Gif my mehr!" Un	je mehr he eet, je	mehr wull he hebben, un säd	1819	1815
17	GRI	Mann, "meine Natur ist ganz anderer Art,	je heißer es ist, desto	mehr frier ich, und der	1819	1815
18	GOE	bedarf weniger der Neuheit, ja vielmehr	je älter sie ist, je	gewohnter man sie ist,	1827	1825
19	MK1	so daß man ebenso sagen könnte:	je parataktischer, desto	epischer (vergleiche	1946	1945
20	Z53	Die Behauptung dieses Raumes wird	desto sicherer gelingen, je	stärker er in das	1953	1955
21	Z53	Haifisch herausgefallen sein." "Je,	je, je",	sagte die Köchin, "Kinder sind	1953	1955
22	Z54	meinen, müßte auf Wahlen drängen, denn	je rascher sie erfolgen, desto	wahrscheinlicher	1954	1955
23	BZK	Entlastung, die man von ihnen erwartet.	je größer eine Stadt, desto	mehr Verkehr saugt	1954	1955
24	Z54	mehr muß ihr Träger Spezialist sein, und	je höher er steigt, desto	weniger Spezialist	1954	1955

Fig. 3: Mit getCOSMAS() erstelltes Spreadsheet mit der Option merge = TRUE: Eine aus mehreren Wörtern bestehende Keyphrase wird in Spalte C („Key“) zusammengeführt.

getNSE

- filename: Der Pfad zur .txt-Datei, die Sie aus NoSketchEngine exportiert haben. Die Hinweise zu *filename* bei *getCOSMAS* oben gelten entsprechend.
- tags: In den allermeisten Korpora (z.B. Wacky oder COW) lassen sich Zusatzinformationen (Tags) wie z.B. Lemma, Wortart etc. exportieren. Diese werden in der Exportdatei durch einen Slash (/) abgetrennt direkt als das einzelne Wort angehängt. Fig. 4 zeigt dies am Beispiel einer Exportdatei mit Lemma-Tags. Ist tag = FALSE, dann bleibt das Keyword bzw. die Keyphrase in diesem Format erhalten. In der Regel will man das jedoch nicht, weil die Keyphrase dann schwer zu entziffern ist, gerade wenn man noch weitere Tags mit exportiert hat. Deshalb ist der Default tag = TRUE: Mit dieser Einstellung werden die Tags in je eine eigene Spalte „ausgelagert“. **Hinweis:** Bei NoSketchEngine kann man Tags entweder für den gesamten Kontext oder nur für die Keyphrase exportieren lassen. Im Gegensatz zu *getCWB()* (s.u.) verfügt getNSE derzeit nicht über eine Funktion, die es erlaubt, die Tags aus dem Kontext zu entfernen. Wenn man die Tags im linken und rechten Kontext nicht braucht, sollte man sie also von Anfang an nicht exportieren, um am Ende eine gut lesbare Konkordanz zu haben.
- convert und new_filename: *netNSE()* bietet außerdem die Möglichkeit, die Konkordanz direkt als tab-separierte Datei zu exportieren. Dafür muss man convert = TRUE setzen und einen Dateinamen für die zu generierende Exportdatei eingeben, z.B. `getNSE(filename = "myfile.txt", filename_new = "mynewfile.txt", convert = TRUE)`

```
[947] " je /je  nach /nach  je /je  nach /nach "
[948] " je /je  verrückter /verrückter  desto /desto  besser /gut "
[949] " je /je  unabhängiger /unabhängig  desto /desto  kritischer /kritisch "
[950] " je /je  früher /früher  desto /desto  besser /gut "
[951] " je /je  Tag /Tag  je /je  Monat /Monat "
[952] " je /je  Monat /Monat  je /je  Jahr /Jahr "
[953] " je /je  länger /lang  je /je  mehr /mehr "
[954] " je /je  eher /eher  desto /desto  besser /gut "
[955] " je /je  schlimmer /schlimm  desto /desto  besser /gut "
[956] " je /je  abgelegener /abgelegen  desto /desto  schöner /schön "
[957] " je /je  mehr /mehr  desto /desto  besser. /besser. "
[958] " je /je  kürzer /kurz  desto /desto  besser. /besser. "
```

Fig. 4: Exportdatei aus NoSketchEngine mit Lemma-Tags.

getWACKY

- filename: Der Pfad zur XML-Datei, die Sie aus NoSketchEngine exportiert haben.
- tags: Wenn die Keyphrase Tags enthält (z.B. Lemma oder Wortart), dann haben Sie hier drei Optionen, mit ihnen umzugehen: tags = "omit" (der Default) entfernt die Tags einfach, tags = "display" zeigt die Tags zusammen mit den Keywords (durch / abgetrennt, so wie in Fig. 4), und tags = "column" lagert sie quasi in eine eigene Spalte aus, so wie tags = TRUE bei *getNSE()*.

getCWB

- filename: Die Exportdatei aus CWB (die sich z.B. in der Terminal-Version von CQP mit cat Last > filename.txt exportieren lässt; zum Einstieg in CQP empfehle ich neben den teils etwas technischen Handbüchern auf http://cwb.sourceforge.net/ auch die Tutorials von Susanne Flach unter https://userpage.fu-berlin.de/~flach/corpling/workshops-materialien/) oder aus CQP*web*.
- dt: Wenn TRUE (der Default), dann werden die Resultate als *data.table* ausgegeben. Um die Verarbeitung großer Dateien zu beschleunigen, stützt sich *getCWB* auf das R-Paket *data.table* (Dowle & Srinivasan 2017). Data.tables erfordern eine etwas andere Syntax als normale Dataframes, die man z.B. mit Hilfe dieses CheatSheets (zuletzt abgerufen November 2017) erlernen kann. Wer die „normale“ Syntax bevorzugt, kann dt = FALSE benutzen und bekommt die Ergebnisse als ganz normalen Dataframe.

Literatur

Dowle, Matt & Arun Srinivasan. 2017. *data.table: Extension of data.frame*. https://CRAN.R-project.org/package=data.table