Hélène Vinckel-Roisin
(Sorbonne Université, Paris)

Freitag, 30. August 2019, 11.00-12.15

Das *Deutsche Referenzkorpus*: Präsentation und Verwendungsmöglichkeiten

Das am Leibniz-Institut für deutsche Sprache (IDS, Mannheim) seit 1967 aufgebaute *Deutsche Referenzkorpus* bzw., seit 2004, *DeReKo* dient der Germanistik als empirische Grundlage für die Erforschung der deutschen Gegenwartssprache. Ziel des Workshops ist es, die Relevanz dieses der Forschergemeinschaft kostenlos zur Verfügung gestellten schriftsprachlichen Korpus für linguistische Fragestellungen aufzuzeigen.

1. Zum Einstieg: Was sind (linguistische) Korpora?

- Korpus – Definition

Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. (Lemnitzer/Zinsmeister 2006, 7)

- Linguistische Korpora

Wenn wir von linguistischen Korpora sprechen, dann handelt es sich um Textsammlungen mit kompletten Texten oder zumindest mit sehr großen Textausschnitten. Außerdem sollten linguistische Korpora meist
- repräsentativ,
- durch Metadaten erschlossen und
- linguistisch annotiert sein. (Lemnitzer/Zinsmeister 2006, 40)

- Relevanz und Nutzen von Korpora

Je größer ein Korpus, desto zuverlässiger können Schlussfolgerungen über seltene und diversifizierte Ereignisse gezogen werden. (Lüngen/Kupietz 2014, 25)

Je größer die Datenmenge und je besser die Art und Weise, wie die Stichprobe gesammelt wurde, desto eher liegt der Schluss nahe, dass sich die Aussagen auf das Querschnittsverhalten der Sprachgemeinschaft beziehen lassen. (Perkuhn et al. 2015, 182)

Das Deutsche Referenzkorpus (DeReKo)

- ✓ ein hervorragendes, frei zugängliches, für Ihre Fragestellungen geeignetes Korpus
- ✓ eine nutzbare Textsammlung zur Erstellung Ihres eigenen Korpus zwecks Beantwortung linguistischer Fragestellungen – quantitative und qualitative Untersuchungen
- ✓ morphosyntaktisch annotierte Korpora[1]; teilweise auch nach Wortarten annotiert (vgl. unten)
- ✓ ein Mittel / ein dem Linguisten zur Verfügung stehendes Instrument, um die Frage zu beantworten: Wie findet man sprachliche Phänomene in großen Mengen von Sprachdaten?

2. Das Deutsche Referenzkorpus (DeReKo)

2.1. Anmeldung und Überlick über die verschiedenen Menüpunkte (im Wegweiser)

Zugang zum *DeReKo* (über die Version 2.2.2 der webbasierten Benutzeroberfläche von COSMAS II) über: https://cosmas2.ids-mannheim.de/cosmas2-web.

- sich anmelden/sich abmelden
- Online-Hilfe
- größte zentrale Korpussammlung Deutschlands – Online-Anfrage mit Cosmas II (= das Korpusrecherchesystem)
- Untermenü „Optionen" – vordefiniert, aber man kann selber für seine eigene Korpusrecherche die Kriterien bestimmen/auswählen.

2.2. Das Menü „Archiv/Korpora" im Einzelnen

- Archiv der geschriebenen Sprache (W-W4) (das Hauptarchiv)
- W-ÜBRIG – Archiv der aussortierten geschriebenen Korpora
- PARONYM – Archiv des Paronym-Projekts des IDS
- WP – Archiv der Wikipedia-Artikel und -Diskussionen (2013/2015/2017)
- WPE – Archiv der englischsprachigen Wikipedia-Artikel und -Diskussionen (2015)
- WP-FS – Archiv der fremdsprachigen Wikipedia-Artikel und -Diskussionen (2015)
- TAGGED – C – Archiv morphosyntaktisch annotierter Korpora (CONNEXOR-*Tagset*) (1994-2009)
- TAGGED – C2 – Archiv morphosyntaktisch annotierter Korpora (CONNEXOR-*Tagset*) (2010-2014)
- TAGGED – T – Archiv morphosyntaktisch annotierter Korpora (TreeTagger-*Tagset*) (1994-2009)
- TAGGED – T2 – Archiv morphosyntaktisch annotierter Korpora (TreeTagger-*Tagset*) (2010-2014)
- TAGGED – M – Archiv morphosyntaktisch annotierter Korpora (MECOLB-*Tagset*)
- HIST – Archiv der historischen Korpora
- GFDS – Kartei der Gesellschaft für deutsche Sprache
- WK – PH – Archiv der phasengegliederten Wendekorpora
- EuroGr@mm–WK (12 Rubriken) – Wikipedia-Artikel + Wikipedia-Diskussionen (deutsch – französisch – ungarisch – italienisch – polnisch – norwegisch)

Die Wikipedia-Korpora: Kurzdarstellung

Aufbereitung und Annotation der Wikipedia-Texte auf die 2. Phase des europäischen Forschungsprojekts *EuroGr@mm/ProGr@mm kontrastiv* zurückgehend (2009-2012)

[1] Zu den Annotationen im DeReKo (die hier weitgehend unberücksichtigt bleiben) vgl. die Informationen und Einzelheiten unter: http://www.ids-mannheim.de/cosmas2/projekt/referenz/annotationen.html.

Vorteil von Wikipedia-Texten: stehen für viele Kontrastsprachen zur Verfügung (= Vergleichskorpora[2]); leicht zugänglich; keinen Urheberrechten unterliegend.
2 Subtextsorten: Wikipedia-Artikel und -Diskussionen, die in ihren sprachlichen Charakteristika sehr unterschiedlich sind.

*Während die **Wikipedia-Artikel** den allgemeinen Merkmalen geschriebener Sprache entsprechen, formelle Charakter aufweisen und in der Regel normgerecht verfasst sind, finden sich bei den **Wikipedia-Diskussionen** Merkmale verschriftlicher mündlicher Sprache, außerdem weisen die Texte teilweise normabweichenden Gebrauch im Bereich der Orthografie und der Syntax auf. Gerade diese Bipolarität der beiden Subtextsorten stellt eine gute Basis dar, sollen synchrone Variationen im Sprachgebrauch ausfindig gemacht werden. (Dalmas/Fabricius-Hansen/Schwinn 2016, 5)*

nach Wortarten annotierte Wikipedia-Teilkorpora: Die deutsche, französische und italienische Version der Wikipedia-Texte wurden mit dem Stuttgarter Tree-Tagger getaggt. Die polnische Wikipedia-Version wurde mit dem Morfeusz-Tagger bearbeitet. Die ungarische (stark) reduzierte Version konnte mit Hilfe eines Teams des *Ungarischen National-Korpus* annotiert werden und schließlich die norwegische Version mit dem Oslo-Bergen-Tagger. So entstanden folgende nach Wortarten annotierte Wikipedia-Teilkorpora:

Übersicht aus Dalmas/Fabricius-Hansen/Schwinn (2016, 6)

Sprache	Wikipedia-Artikel Anzahl der Wortformen	Wikipedia-Diskussionen Anzahl der Wortformen
Deutsch	551.090.404	246.028.026
Französisch	526.944.992	101.893.579
Italienisch	329.063.792	42.171.984
Norwegisch	70.377.449	1.400.276
Polnisch	190.046.721	15.470.942
Ungarisch (Stichprobe)	11.285.752	27.810

- Was bedeutet „nach Wortarten annotiert“? Vgl. den Operator MORPH ASSIST
  Möglichkeit, nach einer bestimmten Wortart zu suchen, z.B. je nach linguistischer Fragestellung: Sammlung von „Partizipien II (voll)“

2.3. Das Menü „Korpusverwaltung“ (Korpusauswahl)

Untermenü (mit 3 Funktionalitäten - Korpusverwaltung) :

Menüpunkt	**Bedeutung**
vordefiniert	Laden eines vordefinierten virtuellen Korpus
benutzerdefiniert	Laden und Löschen eines benutzerdefinierten virtuellen Korpus
geladen	Korpusverwaltung aller geladenen Korpora

(Mehr Informationen und Erläuterungen über die Korpora in der *Online-Hilfe*-Rubrik unter folgendem Hinweis auf der Webseite: „Informationen zu den Korpora finden Sie unter Organisation des Textmaterials.“)

Jedes Korpus besteht aus mehreren Texten/Dokumenten. Anfang 2018 fanden sich in COSMAS II 110881 Dokumente bzw. 212,9 Mio. Texte.

Im Laufe der letzten fünf Jahre hat das IDS schwerpunktmäßig in den folgenden Bereichen neue Textdaten akquiriert (Lüngen 2017, 164):

[2] **Vergleichskorpora** enthalten „Texte mehrerer Sprachen S_1 … S_n zu vergleichbaren Diskursbereichen, die aber keine Übersetzungen voneinander sind.“ (Lemnitzer/Zinsmeister 2006, 104)

*Presse: Die Archive großer überregionaler Tageszeitungen (*Süddeutsche Zeitung, Die Zeit, Der Spiegel*) wurden für DeReKo erschlossen, und durch eine Kooperation mit einem kommerziellen News-Datenbank-Provider wurden zahlreiche weitere regionale und überregionale Pressequellen akquiriert (rund 20 Milliarden Tokens). Dadurch ist der deutsche Sprachraum mit Pressequellen mittlerweile recht gleichmäßig abgedeckt.*

Konzeptionnelle Schriftlichkeit bei medialer Mündlichkeit: Das PolMine-Korpus deutscher Plenar-debattenprotokolle (Uni Duisburg-Essen) und das German Political Speeches-Korpus wurden integriert mit insg. 316 Millionen Token.

Belletristik: Als Spenden von Verlagen wurden seit 2011 neu eingeworbene vollständige Buchtexte im Umfang von circa 8,5 Millionen Tokens in DeReKo integriert. Die vergleichsweise geringe Tokenanzahl dieses Genres erklärt sich dadurch, dass die Akquisition von Buchpublikationen am aufwändigsten ist, sowohl was Verhandlungen mit den Rechte-Inhabern als auch die technischen Anforderungen an die Korpusaufbereitung betrifft.

Konzeptionelle Mündlichkeit bei medialer Schriftlichkeit: Internetbasierte Kommunikation

***a. Wikipedia-Korpora.** In einem Turnus von zwei Jahren wird die jeweils aktuelle deutschsprachige Wikipedia als Korpus in DeReKo zur Verfügung gestellt. Die Konvertierung von 2015 umfasst sämtliche Artikel, alle Artikel-Dikussionen und erstmals auch alle Nutzer-Diskussionen, mit einem Gesamtumfang von 1.378 Milliarden Tokens.*

***b. Ein Usenet-News-Korpus** mit den Texten aller deutschen Newsgruppen seit 2013 wurde für DeReKo aufgebaut. Dieses Korpus wird in Zukunft sowohl aktualisierend als auch mit weiter zurückliegenden Jahrgängen erweitert.*

alle sechs Monate: neue, aktualisierte Ausgabe von DeReKo (Lüngen 2017, 165)

2.4. Forschungsfrage und Eingabe der Suchanfrage

2.4.1. Beispiele für linguistische Forschungsfragen (auf die im Workshop eingegangen wird!)

Beispiele für mögliche linguistische Fragestellungen (aus morphologischer, lexikalischer, syntaktischer, orthografischer Sicht) – die den Ausgangspunkt für die praktische Umsetzung Ihrerseits im Laufe des Workshops bilden ☺

1. Unklarheiten bei Markenbezeichnungen wie auch Fremdwörtern: *die* oder *das Nutella*? *downloadet* oder *gedownloadet*? Welche Variante setzt sich im Sprachgebrauch durch?[3]
2. Welche Wortverbindungen fallen Ihnen z.B. zu *Faden* oder zu *Kopf* oder zu *Willkommenskultur* ein, welche Assoziationen?
3. Suche nach Kollokationsmustern: Beispiele wie *lang* + *Haar* = ist *lange Haare* eine Kollokation? Sind beide Sätze identisch: *Er hat Haare, die lang sind* vs. *Er hat lange Haare*?
4. In welchen Kontexten kommen die Quantifikatoren *ein paar* und *einige* vor? Sind bestimmte Kookkurrenzprofile identifizierbar? Sind diese Lexeme komplementär oder schließen sie einander aus? Vergleichbare Fragestellung mit den Adjektiven *ähnlich* vs. *gleich*? Im Fokus stehen Aspekte von Quasi-Synonymie.
5. Zu einem Aktualitätsthema und einem Neologismus: Wie wird der Ausdruck/die Kollokation *les gilets jaunes* im Deutschen wiedergegeben? *Die Gelbwesten* (als Kompositum) oder die *gelben Westen*? Welches kommt tendenziell in der Pressesprache seit 2018 am häufigsten vor? Lassen sich Kontextunterschiede für die eine oder die andere Variante feststellen? (interessantes und spannendes Thema für eine kontrastiv angelegte MA-Arbeit!)

[3] Vgl. hierzu die Rubrik „Grammatik in Fragen und Antworten“ vom IDS, die auf DeReKo beruht. https://grammis.ids-mannheim.de/fragen

6. Satzanfänge im Sprachvergleich – siehe das *EuroGr@mm-Projekt* zu *Variation am Satzanfang* und die korpuslinguistisch orientierten Untersuchungen anhand der einzelsprachlichen Wikipedia-Teilkorpora (Dalmas/Fabricius-Hansen/Schwinn (Hgg.) 2016; Augustin 2016).
7. Befindet sich das als Nomen fungierende Lexem *Hauptsache* im linken Außenfeld (/Vorvorfeld) im V2-Satz im heutigen Sprachgebrauch auf dem Weg zur Grammatikalisierung? Hauptsache als Diskursmarker/Fokusmarker?
8. Weist z.B. in politischen Reden, Plenarprotokollen oder anderen Textsorten bei offiziellen Anlässen (Danksagungen/ritualisierter Sprechakt) die Wendung *möchte/n..... (jdem) herzlich danken **für** (+NP/PP)* auf eine nicht nur lexikalische, sondern auch syntaktisch-lineare/syntagmatische Festigkeit hin, die sich in der Nachfeldposition der PP *für (+NP/PP)* niederschlägt? Hypothese eines hohen Grades an Routine bzw. Konventionalisierung, der sich in der markierten Linearisierungsabfolge manifestiert.
9. Gibt es Häufigkeitsunterschiede, die mit *nämlich* eingeleitete Zusätze im rechten Außenfeld kennzeichnen, je nach untersuchter Textsorte (Pressetexte aus der *Süddeutschen Zeitung vs.* Goethes Werke *vs.* Wikipedia-Diskussionen)? Wenn ja, wo wird am meisten im Nachhinein reformuliert/präzisiert? Und in welchen Kontexten (Vor- und Folgetext betreffend)?
10. Häufigkeit von mit *und zwar* eingeleiteten Zusätzen in Plenarprotokollen *vs.* im geschriebenen Deutsch (z.B. *Süddeutsche Zeitung*)? Gibt es präferierte Nomen/NPs, Adjektive etc., die an der rechten Satzperipherie gemäß der Sprecherintention vorkommen zwecks bestimmter zu identifizierender Effekte in der analysierten Textsorte (= bestimmte lexikalische Assoziationen mit *und zwar* an der rechten Satzperipherie?)

2.4.2. Allgemeines zu Suchanfragen und kurzes Tutorial[4]

Beispiel-Suchanfrage 1: Gelbwesten (im Korpus sz-Süddeutsche Zeitung (1992-2018)) – siehe Fragestellung 5

Noch etwas, worauf Sie vor der endgültigen Suchanfrage aufmerksam sein sollten => Fenster mit Wortformliste(n):
Die Wortformliste zur gezielten/ausgewählten Suchanfrage enthält im aktiven Korpus generell mehrere Einträge. Wählen Sie alle Zeilen in der Wortformliste, die für Ihre Suche relevant sind/in Frage kommen, an und wählen Sie die anderen Zeilen wieder ab!

Ein Wort zu **den Ergebnissen der Korpusrecherche** und deren Formaten im DeReKo:
- KWIC-Format – Übersicht über alle gefundenen Treffer
- Volltext-Format – unter Einbeziehung des Kontexts vor und nach dem Suchbegriff – ermöglicht, sich einen Überblick über die Kontexteinbettungsbedingungen (und ggf. kleine satzübergreifende Analyse) zu verschaffen. Sehr nützlich und sinnvoll – großer Vorteil von DeReKo!

Beispiel-Suchanfrage 2: Nutella - siehe Fragestellung 1

Was können Korpusbefunde aussagen? Das *Nutella*-Beispiel in DeReKo (Perkuhn/Keibel/Kupietz 2012, 72-73) – Häufigkeit und Relevanz[5]

Untersuchung, die Perkuhn/Keibel/Kupietz 2012 vorgenommen haben (72-73):

> *Mit welchem Anteil am Gesamtvorkommen wird das Wort „Nutella" mit den jeglichen Artikeln verwendet?*
>
> *Bei 1069 Vorkommen insgesamt sind es nur 42, genau genommen sogar nur 28 Textstellen, in denen unmittelbar vor dem Wort „Nutella" ein Artikel verwendet wird. Ist dies nicht die viel gewichtigere Aussage, dass es eine starke Präferenz gibt, das Wort ohne Artikel zu verwenden? (Perkuhn/Keibel/Kupietz 2012, 73)*

[4] Genaueres unter der Rubrik „Online-Hilfe zu Cosmas II – Suchanfrage"
http://www.ids-mannheim.de/cosmas2/web-app/hilfe/suchanfrage/
[5] Vgl. in Ergänzung hierzu den Beitrag von Donalies (2008).

Beispiel-Suchanfrage 3: und zwar (im Korpus *pp-Plenarprotokolle 1*) – siehe Fragestellung 10

Hinweis – Wenn logische Operatoren als Suchbegriffe fungieren – *und, oder, nicht, and, not*, etc. ➔ diese in Anführungszeichen setzen!

Beispiel-Suchanfrage 4 – mit Verwendung des Platzhalters * (beliebig viele oder gar kein Zeichen)

Mond* = Suche nach allen Wortformen – das Wort selbst, mögliche Flexionsformen oder Kompositum mit dem vorgegebenen Wortanfang

Liste der Platzhalteroperatoren *, ?, +
Operator(en) und Wort bzw. Wortsegment(e) bilden das Suchmuster.
Diese Operatoren ermöglichen die Suche nach Wörtern und Wortsegmenten mit ergänzenden Zeichen(ketten). Das Wort (bzw. das Wortsegment) als solches ist ebenfalls gleichzeitig Suchbegriff.
Diese Operatoren sind "Platzhalter" für beliebige Zeichen:
***** steht für 0 bis n Zeichen (keines bis "unendlich" viele),
? steht für genau ein Zeichen,
+ steht für 0 oder 1 Zeichen (höchstens eines).
Diese Operatoren können am Anfang oder am Ende von Wörtern bzw. Wortsegmenten stehen bzw. von Wortsegmenten umgeben sein. Sie sind ohne Leerzeichen mit dem Wort bzw. Wortsegment in das Suchanfragefenster zu schreiben.

- ***Sie sind dran* (Aufgabe 1)!** *Nachdem Sie schon nach dem Kompositum „Gelbwesten" im Korpus* sz-Süddeutsche Zeitung *gesucht haben, möchten Sie die Ergebnisse mit der Adj-Nomen-Verbindung „gelbe Westen" vergleichen. Welche Suchanfrage brauchen Sie, um soviele Treffer wie möglich zu erreichen (und somit zu aussagekräftigen Ergebnissen zu kommen)?*

Beispiel-Suchanfrage 5 - Suche an einer bestimmten Satzposition

Hauptsache im linken Außenfeld – verbunden mit der Fragestellung 7 (vgl. oben) - Hypothese eines Grammatikalisierungsprozesses im heutigen Sprachgebrauch

Hinweis aus der Online-Hilfe

Es gelten hierzu folgende Formeln:

Suchanfrage	Erläuterung	Häufigkeit
wegen #IN(**L**) <s>	*wegen* am Satzanfang	202.206
wegen #IN(**R**) <s>	*wegen* am Satzende	11.095
wegen #IN(**F**) <s>	*wegen* von Satzanfang bis Satzende	50
wegen #IN(**N**) <s>	*wegen* weder am Satzanfang noch -ende	1.455.943
	Summe von L, R, F und N	1.669.294
wegen #IN <s>	keine Spezifizierung, d.h. alle Fälle zusammen	1.669.294

Suchanfrage: **Hauptsache #IN(L) <s>**

- ***Sie sind dran* (Aufgabe 2)!** *Suchen Sie im Teilkorpus* Wiki_DE-öffentlich - öffentliche deutschsprachige Wikipedia-Artikel *nach* Hauptsache *am Satzanfang? Zu welchen Ergebnissen kommen Sie?*

2.5. Exportieren der Forschungsergebnisse

Beispiel von *Hauptsache* im linken Außenfeld in den Wikipedia-Disskussionen (deutsch) (vgl. Fragestellung 7)

Hypothese eines Grammatikalisierungsprozesses: Ergebnisse exportieren – Relevanz für weiterführende Untersuchungen mit dem Ziel, erste Antworten quantitativ und qualitativ auf die linguistische Forschungsfrage zu liefern (vgl. Stichprobe im Anhang am Ende des Papers).

- 763 Treffer – 2005 bis 2011 – ggf. die nicht-relevanten Treffer manuell entfernen (z.B. *Hauptsache ist, dass...*)
- Eine Auswahl von KWIC-Zeilen bzw. Belege im Fenster mit der KWIC- und Belegdarstellung markieren (= die aus Ihrer Sicht relevanten Belege) – ggf. wenn alle Treffer relevant sind, die Option „Nur ausgewählte Treffer" in den Exportkriterien ausschalten!
- Kriterien für das Exportieren der Ergebnisse vordefinieren: Name der Datei, Schriftgröße, Format etc.
- Anschließend: Exportdatei herunterladen!

➢ ***Sie sind dran*** **(Aufgabe 3)!** *Ausgehend von den bisherigen praktischen Übungen/Suchanfragen exportieren Sie die Ergebnisse und speichern Sie sie auf Ihrem Desktop (RTF-Format). Bitte achten Sie auf die auszuwählendenbe nutzerdefinierten Kriterien und bestimmen Sie (unter Berücksichtigung der Forschungsfrage), was für die geplante korpuslinguistische Untersuchung ggf. von Relevanz sein kann (Anzahl der Sätze davor und/oder danach etc.)*

2.6. Kookkurrenzanalyse[6]: ein Überblick

2.6.1. Was sind Kookkurrenzen? Was leistet die Kookkurrenzanalyse?

Im Zusammenhang mit den Fragestellungen 3 und 4:

- lang + Haar = ist *lange Haare* eine Kollokation?
- Sind die Quantifikatoren *einige* und *ein paar* synonym? Verbinden sie sich jeweils präferenziell mit bestimmten Substantiva? Gibt es für jeden – je nach Textsorte – ein präferiertes Assoziationsmuster?

Kookkurrenz und Kollokation

> *Das Phänomen, dass Wörter nicht nur sich selbst beim Verstehen oder Erzeugen eines Satzes (oder Textes) einbringen, sondern auch Einfluss beim Verstehen (und bei der Wahl) der Wörter in ihrer Umgebung haben, ist nicht unbekannt. Es wird seit langem unter verschiedenen Schlagworten diskutiert – z.B.* lexikalische Solidaritäten *(Coseriu 1978). […] Im Standardszenario dreht sich alles um das lexikalische Material, im Mittelpunkt steht der Gebrauch eines bestimmten Wortes (im Folgenden: Bezugswort; als Wortform oder Grundform, je nach Fragestellung). Ausgehend von diesem Wort wird der Ausschnitt definiert, der für die Auswertung der Kookkurrenz maßgeblich ist. Dazu muss das Analyseverfahren quasi in die Texte hineinschauen, in denen das Wort vorkommt. (Perkuhn/Keibel/Kupietz 2012, 110)*

Definition des Kontexts: wie viele Wörter links? wie viele Wörter rechts? Satzgrenzen beachten?

im Vorfeld festlegen,

> *wie viele Wörter davor und danach für die Fragestellung berücksichtigt werden sollen – der zu untersuchende Kontext wird definiert. Dabei ist auch zu hinterfragen, ob das zu untersuchende Phänomen an Satzgrenzen Halt macht oder es womöglich erst zum Vorschein kommt, wenn die Umgebung eines Wortes über Satzgrenzen hinweg betrachtet wird. (Perkuhn/Keibel/Kupietz 2012, 117)*

Vorteile

- eine korpusanalytische Methode zur Sichtbarmachung sprachlicher Strukturen/präferierter Verbindungen/Assoziationen im Sprachgebrauch
- Aussagekraft bzgl. Grades der lexikalischen Festigkeit/Fixiertheit – Frequenz/Häufigkeitskriterium
- Beschreibung usueller Wortverbindungen auf einer umfassenden empirischen Basis
- praktischer Nutzen: In der Lexikografie etwa kann eine Kookkurrenzanalyse helfen, die Lesarten eines Wortes zu bestimmen, die in einem Wörterbuch dargestellt werden sollen.

[6] Ein ausführlicheres Tutorial zu der Kookkurrenzanalyse finden Sie hier:
Perkuhn, Rainer/ Belica, Cyril: „Eine kurze Einführung in die Kookkurrenzanalyse und syntagmatische Muster". Institut für Deutsche Sprache, Mannheim. 2004. http://www1.ids-mannheim.de/kl/misc/tutorial.html

Ausgangspunkt für vertiefende Untersuchungen

- z.B. besonders geeignet für Vergleiche von Quasi-Synonymen

  Unter Synonymie versteht man in der Linguistik eine lexikalisch-semantische Relation zwischen zwei Wörtern, die die gleiche Bedeutung haben. Da sich diese idealisierte Vorstellung im tatsächlichen Sprachgebrauch nicht zeigt, verwenden wir lieber den abgeschwächten Terminus Quasi-Synonymie, *der bereits andeutet, dass sich Wörter nur bis zu einem gewissen Grad oder nicht in bestimmten Aspekten entsprechen. (Perkuhn/Keibel/Kupietz 2012, 136)*

Vgl. hierzu weiter unten **2.6.2**. – Vergleich zw. *einige* und *ein paar*; vgl. zudem die Dissertation von Cécile Delettres (2015) (Untersuchung von Quantifikatoren wie *einige* oder *mehrere* und ihren Entsprechungen im Französischen – basierend auf DeReKo fürs Deutsche); vgl. u.a. auch Steyer (2004)

- Projekt in der Lexik-Abteilung, das auf DeReKo basiert und sich der Kookurrenzanalyse bedient

  PREPCON - Präposition-Nomen-Verbindungen im Kontext
  http://uwv.ids-mannheim.de/prepcon/index.html

  Im Mittelpunkt des Projekts steht ein in der Phraseologie bislang eher vernachlässigter Typ: Präposition-Nomen-Verbindungen mit rekurrenter Nullstelle, im Kern binäre verfestigte präpositionale Wortverbindungen mit Lexemstatus, z.B. *nach Belieben; mit Genugtuung; unter Tränen; nach Jahren; für Sekunden; vor Ort*.

  Beispiel-PREPCON Online temporal – Modul 2 (temporale Präposition-Nomen-Verbindungen) (nach Jahren ; für Sekunden etc.)
  http://uwv.ids-mannheim.de/prepcon/modul2/index.html

  Inventar temporaler Präposition-Nomen-Verbindungen
  http://uwv.ids-mannheim.de/prepcon/modul2/inventar/temporalangaben.html

2.6.2. Beispiele für Kookkurrenzanfragen und Durchführung einer Kookkurrenzanalyse

- Einstellung verschiedener Parameter – hier Default-Fall/Standardfall (vgl. Eingabemaske für Parameter)[7]

Beispiel 5 (verbunden mit Fragestellung 4)

5a) Ausgehend vom Bezugswort „einige“ eine Kookkurrenzanalyse im Teilkorpus *Bellitristik des 20. und 21. Jahrhunderts* des Archivs der geschriebenen Sprache durchführen – alle Wortformen (mit Flexionsmorphemen) einbeziehen!

Suchanfrage: einige*
Kookkurrenzanalyse aktivieren und Einstellungen definieren (0 Wörter links; 5 Wörter rechts)

Liste mit den häufigsten Verbindungen *einige* + Nomen im Teilkorpus *Bellitristik des 20. und 21. Jahrhunderts* erstellen (= Kookkurrenzprofil).

5b) Dasselbe erneut ausführen: Kookkurrenzanalyse mit dem Bezugswort *ein paar* im Teilkorpus *Bellitristik des 20. und 21. Jahrhunderts* des Archivs der geschriebenen Sprache durchführen!

Suchanfrage: ein paar (/+w1)

[7] Für eine Kookkurrenzanfrage können verschiedene Parameter eingestellt werden. Einige der Parameter legen den zu analysierenden Kontext fest, andere Parameter steuern die Vorgehensweise bei der Analyse. Schließlich gibt es Parameter, die sich auf die Darstellung der Ergebnisse auswirken.

- **Sie sind dran (Aufgabe 4)!** *Die häufigsten Nomen-Verbindungen mit dem Quantifikator* mehrere *im Teilkorpus* Bellitristik des 20. und 21. Jahrhunderts *anhand einer Kookkurrenzanalyse aufzeigen!*
  Vergleichen Sie die Ergebnisse der Kookurrenzanalyse mit den häufigsten Nomen-Verbindungen mit mehrere *im Teilkorpus* Fußball Spielberichte (kicker.de)!

Weitere Übungen für „zu Hause“:

- Sind die Adjektive *leise* und *still* quasi-synonym? Gibt es hierfür statistisch fundierte Evidenz? Was zeigt eine Kookkurrenzanalyse in einem vordefinierten Teilkorpus?
- Vergleichbare Fragestellung mit den Adjektiven *hervorragend / ausgezeichnet / exzellent*, die z.B. in dt-frz. Wörterbüchern bekanntlich als Synonyme im Eintrag *excellent* behandelt werden und häufig den DaF-Lernenden im Übersetzungskurs Schwierigkeiten bereiten! Was zeigt eine auf DeReKo basierende Suche?

3. Zusätzliche Aufgaben im Zusammenhang mit dem Kurs über die Stellungsfelder

- **Sie sind immer noch dran! (Aufgabe 5)** *Zwei Fliegen mit einer Klappe schlagen – eine einzige Suchanfrage, um statistisch/quantitativ fundierte Antworten auf die Fragestellung 8 (vgl. oben) zu bekommen –* **möchte .... (jdem) herzlich danken für (+ NP/PP).**

  Gibt es neben dem hohen Grad an lexikalischer Festigkeit korpuslinguistische empirische Evidenz für eine hohe Fixiertheit, was die ***Nachfeldposition*** *der PP* für (+NP/PP) *in Dankesakten betrifft? Wie lautet die Suchanfrage?*
  Zu welchen Ergebnissen kommen Sie im Teilkorpus Teilkorpus rei - Reden und Interviews, Januar 2002 - Dezember 2006 [2]*?*

- **Aufgabe 6 – im Zusammenhang mit der Fragestellung 9: Textsortenspezifizität von „nämlich X“ im rechten Außenfeld**

  Vergleichen Sie die Vorkommenshäufigkeit von nämlich *im rechten Außenfeld (als lexikalischer Marker eines Zusatzes) in Pressetexten vs. Goethes Werken vs. Wikipedia-Diskussionen! Zusatzübung: Kookkurrenzanalyse und Exportieren der Ergebnisse.*

4. Schlussbetrachtungen

- Frage der Repräsentativität – immer vorsichtig bleiben
- Tipps

  1. Man sollte sich vor der Korpusrecherche darüber im Klaren sein, ob man das Korpus befragt, um die Bestätigung einer Vorannahme zu erhalten oder auch um geeignete Belege zu finden. Ein grundsätzlich anderer Zugang zum Korpus ist, sich von diesem überraschen zu lassen.

  2. Übung macht den Meister: Ausprobieren! Ausprobieren! Ausprobieren!

- Blick in die Zukunft! Korpusanalyseplattform der nächsten Generation – am IDS

  Korap Corpus Analysis Platform = https://korap.ids-mannheim.de

Literatur und weiterführende Links

Augustin, Hagen (2016): Quantitative Untersuchungen zum deutschen Vorfeld und seinen Äquivalenten in sechs verschiedensprachigen Wikipedia-Korpora. In: Dalmas, Martine/Fabricius-Hansen, Cathrine/Schwinn, Horst (Hgg.), *Variation im europäischen Kontrast – Untersuchungen zum Satzanfang im Deutschen, Französischen, Italienischen, Norwegischen, Polnischen und Ungarischen*. Berlin/New York, de Gruyter. 9-52.

Bubenhofer, Noah (2006): *Einführung in die Korpuslinguistik: Praktische Grundlagen und Werkzeuge*. (Elektronische Ressource). Zürich.
https://www.bubenhofer.com/korpuslinguistik/kurs/

Coseriu, Eugenio (1978): Lexikalische Solidaritäten. In: Geckeler, Horst (Hg.), *Strukturelle Bedeutungslehre*. Darmstadt: Wissenschaftliche Buchgesellschaft, 239-253.

Delettres, Cécile (2015): *La sémantique des quantificateurs vagues : étude contrastive allemand-français*. Dissertation. Paris-Sorbonne.

Donalies, Elke (2008). *Der, Die* oder *Das Nutell*a? Zum Genus von Produktnamen. *Sprachreport 8/2004*, 23-25.
https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/2454/file/Donalies-Der_Die_oder_Das_Nutella-2008.pdf

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji (eds.), *Working Papers in Corpus-based Linguistics and Language Education, No. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.
http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=23DEEEE7120DDDCDA69A4C17B3E00D91?doi=10.1.1.475.9714&rep=rep1&type=pdf

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas, 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In: Calzolari, Nicoletta et alii. (Hgg.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki: European Language Resources Association (ELRA), 4353-4360.

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Minegishi, Makoto / Kawaguchi, Yuji (Eds.): *Working Papers in Corpus-based Linguistics and Language Education, No. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.
http://www.lrec-conf.org/proceedings/lrec2018/pdf/737.pdf

Lemnitzer, Lothar/Zinsmeister, Heike, 2006. *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Lüngen, Harald (2017): DeReKo – Das Deutsche Referenzkorpus. *Zeitschrift für germanistische Linguistik*, 45(1), 161–170.
https://doi.org/10.1515/zgl-2017-0008

Lüngen, Harald/Kupietz, Marc (2014): Das Deutsche Referenzkorpus DEREKO im Jubiläumsjahr 2014. *Sprachreport*, 24-28.
https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3039/file/3134Luengen_Kupietz_DEREKO_2014_3.pdf

Perkuhn, Rainer/Belica, Cyril (2004): *Eine kurze Einführung in die Kookkurrenzanalyse und syntagmatische Muster*. Institut für Deutsche Sprache, Mannheim.
http://www1.ids-mannheim.de/kl/misc/tutorial.html

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): *Korpuslinguistik*. Paderborn: Fink.

Perkuhn, Rainer/Belica, Cyril/Keibel, Holger/Kupietz, Marc/Lüngen, Harald (2015): Valenz und Kookurrenz. In: Domínguez Vázquez/ José, Maria/Eichinger, Ludwig M. (Hgg.), *Valenz im Fokus: grammatische und lexikografische Studien; Festschrift für Jacqueline Kubczak*. Mannheim: Institut für Deutsche Sprache, 175-196.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hg.), *Wortverbindungen – mehr oder weniger fest.* Berlin/New York: de Gruyter, 87-116.

Weiß, Christian (2005): *Die thematische Erschließung von Sprachkorpora*. (= OPAL - Online publizierte Arbeiten zur Linguistik 1/2005). Mannheim: Institut für Deutsche Sprache.
https://pub.ids-mannheim.de/laufend/opal/pdf/opal2005-1.pdf

Weitere spezifisch(er)e Fragen zu DeReKo und/oder zu bestimmten Suchanfragen? Dann können Sie jederzeit auch den **Leiter des Programmbereichs Korpuslinguistik am Leibnitz-Institut für Deutsche Sprache, Dr. Marc Kupietz**, kupietz@ids-mannheim.de und / oder **Rainer Perkhun**, perkuhn@ids-mannheim.de, kontaktieren!

Anhang - Auszug

Hauptsache im linken Außenfeld – Suchanfrage (*Hauptsache* am Satzanfang)

COSMAS II-Server, C2API-Version 4.17.1 - 26.07.2019

```
Datum         : Mittwoch, den 7. August 2019, 10:32:16
Archiv        : EuroGr@mm-WPD-DE-DISC - Wikipedia Diskussionen (deutsch - TreeTagger)
Korpus        : Wiki_DE_DISC-öffentlich - öffentliche deutschsprachige Wikipedia-Diskussionen
Archiv-Release: Sonstige IDS-Korpora (2010-II)
Suchanfrage   : IN(L,Min)
                 └Hauptsache
                 └ELEM(S){w=0:0,sp=0:0}
Suchoptionen  : Ei+Ri+Di, Flex
Ergebnis      : 763 Treffer
```

Belege (unsortiert)

```
Anz. Treffer           : 763
Anz. exportierte Belege: 763
Angezeigter Kontext    : 5 Sätze links, 3 Sätze rechts
Kontext umschließt     : gesamten Treffer
```

Hast du nicht die ganze Zeit versucht, uns zu überzeugen, dass es eben kein Allgemeinplatz ist? Aber egal, ich werde mal schauen, was ich für historische Auseinandersetzungen über das AA finde, die den Unterschied zur heutigen pragmatischeren Sicht verdeutlichen. --[[Benutzer:Xario|χario]] 04:16, 3. Jan. 2011 (CET)
Kommt halt drauf an, ob man den Satz auf das Zeug davor bezieht oder einfach so in den Raum stellt. Objektiv finde ich ihn in keinem Fall. **Hauptsache** wir sind alle glücklich, frohes Neues. :D --Chricho 12:21, 3. Jan. 2011 (CET)
zentrale Formulierung
Im Artikel wird das Auswahlaxiom in einer nicht-elementaren Form formuliert mit Hilfe des Funktionsbegriffs, der bekanntlich nicht elementar ist. Das ist deswegen problematisch, weil das Auswahlaxiom eigentlich im üblichen ZF keinen Funktionsbegriff voraussetzt. **(WDD11/A01.50380: Diskussion:Auswahlaxiom, In: Wikipedia - URL:http://de.wikipedia.org/wiki/Diskussion:Auswahlaxiom: Wikipedia, 2011)**

Ich werd den Artikel jetzt verschieben. --::Slomox:: >< 23:49, 20. Mär. 2007 (CET)
Ich denke, im deutschen Sprachkontext können beide Formen als gleichberechtigt behandelt werden. In Schleswig-Holstein ist Apenrade natürlich verbreiteter, aber man soll ruhig die dänische Form verwenden, zumal das in der deutsch-dänischen Zusammenarbeit so üblich ist. **Hauptsache,** der erste Satz nennt Apenrade nicht unter Fernerliefen, sondern fett neben Aabenraa. --Arne List 13:46, 12. Apr. 2009 (CEST)
Die deutsche Minderheit ist heute im Königreich Dänemark als einzige Minderheit mit ihrer sprachlichen und kulturellen Besonderheit entsprechend der Rahmenkonvention des Europarates zum Schutz nationaler Minderheiten und durch die Charta zum Schutz von Regional- und Minderheitensprachen anerkannt. Dies hat zur Folge, dass die deutschen Ortsnamen gleichberechtigt neben dänischen genutzt werden können und sollen, was selbst in von der Kommune herausgegebenen Fremdenführern gemacht wird. **(WDD11/A01.74731: Diskussion:Aabenraa, In: Wikipedia - URL:http://de.wikipedia.org/wiki/Diskussion:Aabenraa: Wikipedia, 2009)**

Einschätzung ist in nämlich in literaturgeschichtlichen Lexikas regelmäßig zu finden. Gibt Besprechungen von Koestler, die sich lang und breit damit befassen, wie er darauf nur gekommen ist. zB Henning Ritter, FAZ, 2006. Ist zu googeln. Aber mancher möchte hier ja wohl unter Beweis, dass er weder von Naturwissenschaft, noch von Literaturgeschichte Ahnung hat, noch überhaupt Koestler gelesen hat. **Hauptsache,** Giros Edit revertieren. Selbst wenn man damit zeigt, dass man nicht mal von der Beschränktheit seines eigenen Wissens eine Ahnung hat. Giro Diskussion 15:10, 13. Jul. 2010 (CEST)
Es handelt sich um einen offenen POV. **(WDD11/A02.31200: Diskussion:Arthur Koestler, In: Wikipedia - URL:http://de.wikipedia.org/wiki/Diskussion:Arthur_Koestler: Wikipedia, 2010)**

Der Amazonas hat ja an manchen Stellen mehr als fünf Kilometer breite. Wenn du da jetzt in einer Schleife von 10 km Durchmesser etwas außermittig fährst, fehlen dir schon ein oder zwei Kilometer, oder hast zuviel, und das bei nur einer Schleife. Gruß -- Rainer Lippert (+/-) 23:50, 23. Mai 2007 (CEST)
Guten Morgen, nach der Bebauung der Stadt Manaus glaube ich das die Bilder viel aelter sind. Mein Haus ist 8 Jahre alt und nicht zu finden. **Hauptsache** wir sind schonmal auf dem gleichem Nenner :-). Beim anschauen dieser Bilder kannst du leider nicht sagen zu welcher Zeit die Bilder gemacht wurden (Hochwasser oder Trockenzeit) da der Fluss von Dezember bis Juni um ca. 3 cm --am Tag steigt und ab Ende Juli bis Ende Oktober um durchschnittlich 12 meter x/Jahr faellt. Es gibt Stellen die sind Ende Juni ueber 20KM breit und Anfang November weniger als 10KM.
Laut telefonischer Auskunft der Capitania (Flussbehoerde und Schiffsueberwachung) wurde die Laenge von 6788 KM als Mittelmass emittelt. Gruesse --Christoph2007 13:44, 24. Mai 2007 (CEST) **(WDD11/A02.47207: Diskussion:Amazonas, In: Wikipedia - URL:http://de.wikipedia.org/wiki/Diskussion:Amazonas: Wikipedia, 2011)**