Originalveröffentlichung in: Mitteilungen des Deutschen Germanistenverbandes Jg. 69 (2022) Nr. 2, S. 142-151.
DOI: *https://doi.org/10.14220/mdge.2022.69.2.142*

POSTPRINT

Rainer Perkuhn

Angebote zu den Korpora der deutschen Gegenwartsschriftsprache am Leibniz-Institut für Deutsche Sprache, Mannheim

Einleitung

Korpora sind – als idealerweise digital verfüg- und auswertbare Sammlungen von Texten – eine wertvolle empirische Grundlage linguistischer Studien. Eigene Korpora aufzubauen ist, je nach Sprachausschnitt, mit unterschiedlichen Herausforderungen verbunden. Zu allen Texten sollten Metadaten zu den Textentstehungsbedingungen (Zeit, Quelle usw.) erhoben werden, um diese als Variablen in Auswertungen einbeziehen zu können. Andere Informationen wie etwa die Themenzugehörigkeit (oder Annotationen auch unterhalb der Textebene) sind auch hilfreich, in vielerlei Hinsicht aber schwieriger pauschal taxonomisch vorzugeben, geschweige denn, operationell zu ermitteln. Jenseits der »materiellen« Verfügbarkeit der Texte und der technischen Aufbereitung sind es das Urheberrecht, vor allem Lizenz- bzw. Nutzungsrechte, sowie ethische Verantwortung und Persönlichkeitsrechte, die beachtet werden müssen, auch um zu gewährleisten, dass die Daten für die Reproduktion der Studien Dritten rechtssicher zugänglich gemacht werden dürfen.

Bevor für ein Vorhaben ein neues Korpus aufgebaut wird, sollte deshalb am besten geprüft werden, ob nicht ein geeignetes bereits zur Verfügung steht. Wenn ein Korpus aufgebaut wird, sollte für eine nachhaltige Aufbewahrung[1] und Zugänglichmachung gesorgt und die Existenz an geeigneter Stelle dokumentiert werden.[2]

Korpora des Leibniz-Instituts für Deutsche Sprache: Das Deutsche Referenzkorpus DeReKo

Für den rechtlich besonders schwierigen Bereich der Gegenwartssprache bietet das Leibniz-Institut für Deutsche Sprache in Mannheim Korpussammlungen an (vgl. Teubert/Belica 2014, Kupietz/Schmidt 2015), für den Bereich der Schriftsprache das Deutsche Referenzkorpus DeReKo.[3] Nach unserem Stand ist dies die weltweit größte, linguistisch motivierte und kontrolliert aufgebaute Sammlung

1 http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_recht.pdf (28.1.2022).
2 https://www.clarin.eu/content/virtual-language-observatory-vlo (28.1.2022).
3 https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (28.1.2022).

von Korpora deutschsprachiger Texte mit einem Gesamtumfang von zzt. über 50 Milliarden Textwörtern (Stand 28. 1. 2022). Neben der mit Abstand am stärksten vertretenen Textsorte »Zeitungen« umfasst der Bestand auch weitere Textsorten wie Belletristik, Publikums- und Fachzeitschriften, Texte der internetbasierten Kommunikation (die komplette deutschsprachige Wikipedia samt Diskussionen, ein Chat-Korpus, Daten aus Usenet-Gruppen, demnächst auch ausgewählte Twitter-Daten) sowie vieles weitere mehr.[4] Auch wenn deren jeweilige Umfänge gering erscheinen, sind die absoluten Zahlen im Vergleich auch international mit anderen Referenz- und Nationalkorpora durchaus wettbewerbsfähig. Wenn für die eigene Studie eine nach eigenen Kriterien ausgewogene Zusammensetzung relevant ist, ermöglichen unsere Recherchesysteme das Arbeiten mit sogenannten *virtuellen Korpora* (s. u.).

Alle Texte liegen in einem einheitlich aufbereiteten Korpusformat[5] vor, mit aussagekräftigen Metadaten und von verschiedenen Werkzeugen mehrfach annotiert (z. B. *part-of-speech* oder Dependenz). Die Daten werden turnusmäßig an die vom IDS angebotenen Recherchesysteme übergeben. Die Nutzung der Daten ist lizenzrechtlich geklärt und abgesichert. Für den Großteil der Daten geben die Vereinbarungen mit den Textgebern allerdings vor, dass sie nicht frei herausgegeben, sondern nur über die besagten Recherchesysteme zugänglich gemacht werden dürfen.

Recherchesysteme

Das IDS unterhält zzt. zwei Recherchesysteme, im Produktionsbetrieb Cosmas II,[6] im Testbetrieb KorAP.[7] Mit beiden kann im Deutschen Referenzkorpus und teilweise in weiteren Korpora über eine Weboberfläche kostenlos recherchiert werden. Cosmas II setzt grundsätzlich, KorAP gestuft nach Lizenz, eine Registrierung voraus, bei der sich die Endnutzer auf eine rein wissenschaftliche und nicht-kommerzielle Nutzung der Daten verpflichten. Für speziellere Auswertungen gibt es weitere Zugangsmöglichkeiten sowie auch weitere Angebote des IDS.

4 https://www.ids-mannheim.de/digspra/kl/projekte/korpora/archiv-1/ (28. 1. 2022).
5 https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell/ (28. 1. 2022).
6 https://cosmas2.ids-mannheim.de/cosmas2-web/ (28. 1. 2022).
7 https://korap.ids-mannheim.de/ (28. 1. 2022).

Cosmas-/KorAP-Bedienoberflächen

Mit Cosmas II kann leider nicht im gesamten Referenzkorpus gleichzeitig recherchiert werden. Aus technischen Gründen musste die Datensammlung auf vier Archive verteilt werden. In weiteren Archiven steht auch Material bereit, das nicht dem Referenzkorpus entstammt. Der erste Schritt einer Cosmas-II-Sitzung erwartet die Auswahl eines dieser Archive, im zweiten Schritt kann entweder dessen Gesamtbestand oder ein Teilkorpus ausgewählt oder (letzteres) als sogenanntes virtuelles Korpus definiert werden. Für den ausgewählten Datenbestand kann dann in der Cosmas-II-Suchanfragesprache ein Rechercheauftrag formuliert werden. Möglich sind Suchen nach Wortformen (sensitiv/insensitiv), nach Lemmata (Flexion plus evtl. komplexere Wortbildung), nach Ausdrücken mit *Wildcards* oder regulären Ausdrücken üblicher Mächtigkeit sowie Kombinationen dieser Einwortausdrücke über ein-/ausschließende Abstandsoperatoren unterschiedlicher Metriken (Wort, Satz, Absatz) sowie über logische Verknüpfungen (und noch vieles mehr).

Über KorAP kann leider noch nicht im gesamten Referenzkorpus recherchiert werden, da bisher nur ein Teil dafür aufbereitet ist. Perspektivisch kann dann gleichzeitig in allen Daten gearbeitet werden. Die Auswahl der Daten über vor- oder selbstdefinierte benannte (Teil-)Korpora wird in absehbarer Zeit möglich sein. KorAP unterstützt aber bereits jetzt die Möglichkeit, virtuelle Korpora als Teil der Suchanfrage über die Metadaten dynamisch zu definieren und zu verwenden.

Beide Systeme bieten als Präsentation des Rechercheergebnisses Konkordanz- und Volltextanzeigen, Cosmas II insbesondere auch verschiedene Sortierungen dieser Darstellungen. Darüber hinaus liefert Cosmas II Ergebnisübersichten, die Nutzer nach verschiedenen Metadaten einstellen können (Zeit, Quelle usw.). Für eine weitere Form der Treffermengenstrukturierung steht unter Cosmas II die Kookkurrenzanalyse zur Verfügung, die nicht nur typische Verwendungsmuster ermittelt, sondern auch in anderen Zusammenhängen wichtige Erkenntnisse ermöglicht.

In beiden Systemen können Sitzungsinformationen und die Treffermengendarstellungen exportiert werden.

Während die Ergebnisübersichten und die Kookkurrenzanalyse nach wie vor die großen Stärken von Cosmas II sind, ist die Erstellung eigener virtueller Korpora hingegen nur begrenzt unterstützt, der Umgang mit Annotationen ist auf kleine Korpora mit einschichtigen Angaben beschränkt.

Diese beiden Punkte sind neben der Skalierbarkeit auf prinzipiell beliebig große Korpora die Stärken von KorAP. Ergebnisübersichten können leider noch nicht angeboten werden.

In KorAP kann mit mehreren Suchanfragesprachen gearbeitet werden, die alle die Definition virtueller Korpora über die Metadaten als Teil der Suchanfrage unterstützen. Mit Poliqarp wird speziell auch eine Anfragesprache angeboten, die

die mehrschichtigen Annotationen von DeReKo handhabbar macht: Die Aussagen verschiedener Werkzeuge können in einer Anfrage kombiniert werden, so z. B. auch um zu evaluieren, ob die Werkzeuge vergleichbare oder sich widersprechende Werte annotiert haben.

Schnittstellen zur Erstellung eigener Abfrage- und Auswertungsskripte

Ein Korpusrecherchesystem ist ein sehr spezielles, komplexes, entsprechend aufwändig zu implementierendes Softwaresystem. Cosmas II wurde weitestgehend allein vom IDS entworfen und implementiert (vgl. Bodmer Mory 2014); KorAP wurde zwar ebenfalls federführend vom IDS entworfen, hat aber eine offene Architektur vorgesehen, bei der die Kernfunktionalität des Systems um Module Dritter erweitert werden kann. Für einen leichten Einstieg in das ›gemeinsame Programmieren‹ bietet KorAP Schnittstellen an, auch um auszuloten, welche Erweiterungen mittelfristig in Form von Standardmodulen eingebunden werden. Diese Offenheit kann sich das System nur deshalb erlauben, da es über ein mächtiges Nutzer- und Lizenzrechtemanagement verfügt. Cosmas II bietet hingegen nur eine intern zugängliche Schnittstelle an.

Für KorAP stehen zzt. Client-Bibliotheken für die Schnittstellen für die beiden Programmiersprachen R[8] und Python[9] zur Verfügung. Die Dokumentationen bieten Beispielskripte an, die leicht abgewandelt werden können, um z. B. die Treffermengen eigener Recherchen statt in Tabellenform als Grafik visualisiert darzustellen – oder z. B. weitere quantitative Auswertungen vorzunehmen.

```
library(RKorAPClient)
library(highcharter)

# Liste der Suchanfragen
query = c("Reaktorunfall oder Reaktorkatastrophe", "Atomausstieg")

# virtuelles Korpus definieren
vc = "textType = /.*[zZ]eit.*/ & availability!=QAO-NC-LOC:ids &
pubDate in"

# Zeitraum festlegen
years = c(1990:2020)
```

8 https://github.com/KorAP/RKorAPClient/ (28. 1. 2022).
9 https://github.com/KorAP/PythonKorAPClient/ (28. 1. 2022).

```
# Verbindung zu KorAP herstellen, Frequenzen über Query erfragen
# mit Highcharter Zeitverlauf mit Konfidenzintervallen zeichnen
new("KorAPConnection", verbose=T) %>%
 frequencyQuery(query, paste(vc, years), ql="cosmas2") %>%
 hc_freq_by_year_ci() %>%
 print()
```

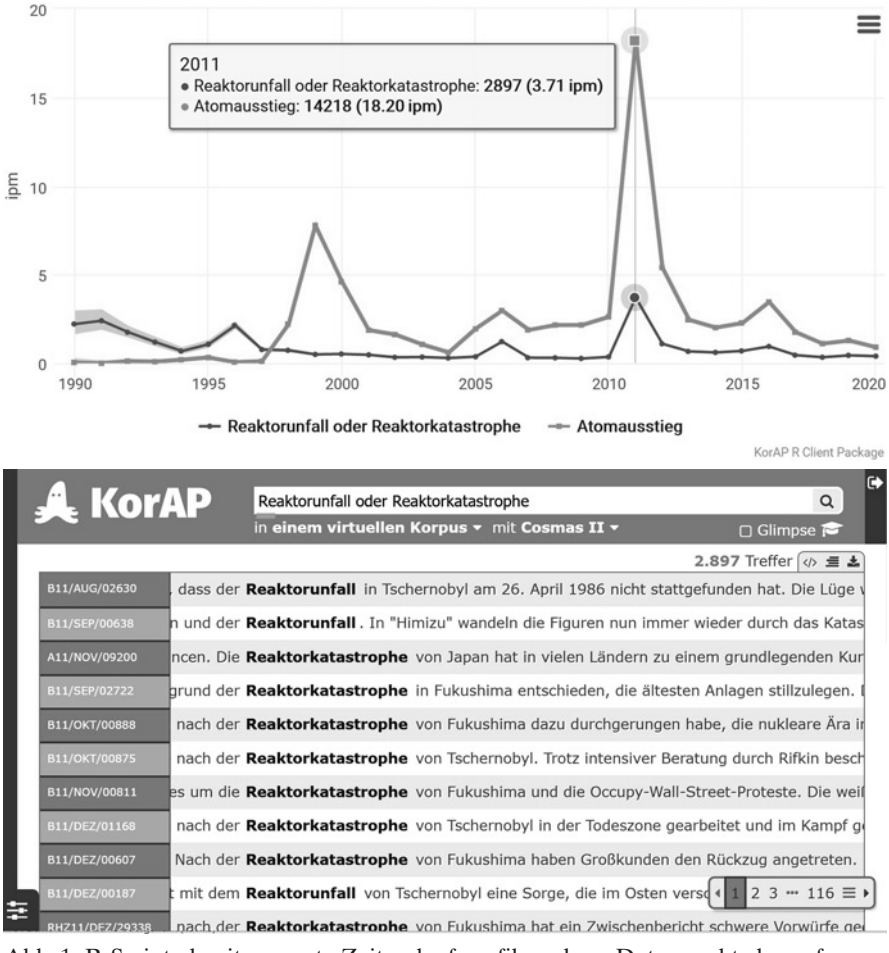

Abb. 1: R-Script, damit erzeugte Zeitverlaufsgrafik und aus Datenpunkt abgerufene KWIC-Ansicht einer KorAP-Anfrage

Durch die in KorAP hinterlegte Mächtigkeit im Umgang mit den Metadaten sind der Vielfalt der Anwendung auf Querschnitts- und Vergleichsanalysen kaum Grenzen gesetzt. Die Verknüpfung von Datenpunkten mit KorAP-Suchanfragen lässt aus interaktiven Grafiken[10] Explorationsinstrumente entstehen.

Tab. 1: Vergleich der beiden Recherchesysteme des IDS

	Cosmas II	KorAP beta
Zugang	nur für registrierte Nutzer	für frei zugängliche Daten auch ohne Registrierung, sonst nur für registrierte Nutzer
Datenbestand	aktuelle Version von DeReKo verteilt auf vier Archive, sowie weitere Korpora, jeweils nur getrennt recherchierbar	zzt. Teilmenge von DeReKo, zukünftig Gesamt- DeReKo und weitere Korpora gemeinsam recherchierbar
virtuelle Korpusdefinition	von Cosmas-Administration prinzipiell textweise, auch gemäß Metadaten (auf Anfrage), nutzerseitig nur über eingeschränkte GUI über Bestandteile der Korpusbezeichner (z. B. Quelle, Zeit)	(zzt. nur:) beliebig als Teil der Suchanfrage über alle Metadaten, zukünftig auch über benannte virtuelle Korpora
Suchanfragesprache(n)	Cosmas II QL	Cosmas II QL, Poliqarp, Annis QL,CQL v1.2, FCSQL
Ergebnisübersichten	nach Metadaten und weiteren vorgegebenen Kriterien	zzt. in Vorbereitung
Exportmöglichkeit	ja	ja
API	nur IDS-intern über CII-script	öffentlich, unterstützt durch KorAP-Client-Bibliotheken für R und Python

10 wie z. B. aus dem Highcharter-Paket (www.highcharts.com (28. 1. 2022)), vgl. Abb. 1.

Tab. 1 *(Fortsetzung)*

	Cosmas II	KorAP beta
eigene Installation/ eigene Daten einspeisen	nein	ja,[11] wenn technische Voraussetzungen gegeben sind (kein Support und keine Betreuung des laufenden Betriebs von Seiten des IDS)

Weitere Angebote

Auswertungen, bei denen eine hohe Rechen- oder Datenlast entsteht, lassen sich nicht über die Recherchesysteme oder deren Schnittstellen umsetzen. Sofern es uns möglich ist, führen wir deshalb ausgewählte Auswertungen selber durch und stellen die Ergebnisse zur Verfügung.

Häufigkeitslisten

Frequenzlisten für beliebige Korpora zu erstellen, ist ein bekanntes Desiderat, das für KorAP auf der mittelfristigen Agenda steht. Cosmas II bietet keine derartige Funktionalität.

Im Arbeitsschwerpunkt DeReWo[12] (Wortlisten zum Deutschen Referenzkorpus) wurden im Laufe der letzten Jahre verschiedene Frequenzlisten mit reflektierten Begriffen zur Lemma-, Wortform- und Zeichenebene erstellt und dokumentiert und werden auf der Webseite zum Download angeboten.

CCDB

Für die explorative Analyse und die vergleichende Auswertung des typischen Kontextverhaltens wurden in der Kookkurrenzdatenbank CCDB[13] die Ergebnisse von über 220.000 Analysen festgehalten. Neben gespeicherten Kookkurrenzprofilen bietet die Plattform die Anwendung verschiedener Methoden, die auf Ähnlichkeiten zwischen Profilen beruhen. Die in der Liste der ähnlichen Profile versteckten, unterschiedlichen Ausprägungen der Ähnlichkeiten lassen sich weiterhin z. B. in einer *self-organizing map* (SOM) kartieren oder für eine Kontrastierung von Wortpaaren (CNS) einsetzen.

11 s. https://github.com/nytud/korap_docker (28. 1. 2022).
12 https://www.ids-mannheim.de/digspra/kl/projekte/methoden/derewo/ (28. 1. 2022).
13 http://corpora.ids-mannheim.de/ccdb/ (28. 1. 2022).

DeReKo-Vecs

DeReKo-Vecs[14] verfolgt einen alternativen Ansatz der Distributionellen Semantik. Dabei werden direkt die Ähnlichkeiten ausgehend von allgemeinen, aber global reduzierten Kontextvektoren mittels neuronaler Netze als *word embeddings* ermittelt und zweidimensional verflacht repräsentiert. Als Kookkurrenzen werden hierbei die an der vom neuronalen Netz gelernten Ähnlichkeitsgewichtung beteiligten Komponenten im Nachhinein rekonstruiert.

Tab. 2: Vergleich der beiden Experimentierplattformen (*open labs*) des PB Korpuslinguistik des IDS

	CCDB	DeReKo-vecs
Datenbestand	virtuelles DeReKo-Teilkorpus (Stand 2007)	wird regelmäßig auf aktuellem Bestand neu berechnet
Referenzstruktur	Kookkurrenzprofile zu Lemmata (ca. 220.000 Bezugswörter)	*word embeddings* zu Token (oder Kombinationen von Token)
Einstieg	über Bezugswörter und Partnerwörter, für Vergleiche ggf. mit Vergleichswort	Token bzw. Tokenkonjunktion (für eine Einheit) oder Tokendisjunktion (für mehrere Einheiten)
Ähnlichkeitsberechnung	über eigenes Maß zum Profilvergleich	entsprechend der *embeddings*
Kartierung	SOM (für einen Eintrag), CNS (für Kontrastierung zweier Einträge)	t-SNE und SOM (für einen Eintrag oder Kombinationen von mehreren Einträgen, entspricht dann Verallgemeinerung von CNS)
weitere Ansichten/Methoden	s. dort	s. dort

Beide Ansätze verfolgen den Grundgedanken, dass sich ihre ›Vektoren‹ (wortbezogen ermittelte Kookkurrenzprofile vs. global reduzierte *embeddings*) sozusagen als räumliche Positionen – und deren »Verwandtschaft« als Nähe – im hochdimensionalen ›Universum der Bedeutungen‹ interpretieren lassen. Der Unterschied zwischen ihnen liegt darin, welche Dimensionen letztendlich zum Tragen kommen und ob die Skalierung der Achsen des angesetzten ›Koordina-

14 http://corpora.ids-mannheim.de/openlab/derekovecs/ (28.1.2022).

tensystems‹ für alle Wörter und Wortvergleiche fest vorgegeben oder situativ ausgeprägt ist.

DeReKo-Bubbles

Genau genommen variiert das ›Bedeutungsuniversum‹ und somit auch die Ausrichtung und Kalibrierung der Koordinatensystems je nach Sprachausschnitt, z. B. je nach Zeitabschnitt oder thematischer Domäne. Eine Vergleichbarkeit kann quasi nur durch einen Fixpunkt (ein initiales oder alternativ sozusagen ein Meta-Universum) hergestellt werden. Darauf aufbauend lassen sich dann z. B. die Positionen desselben Wortes in den verschiedenen Räumen vergleichen und mit weiteren Variablen verknüpfen. In mehreren, auf DeReKo-Vecs aufbauende Studien werden – neben der üblichen räumlichen Anordnung – die weiteren Informationen als Bewegungsbahn sowie als Größe und Einfärbung von Kreisen, als sogenannte DeReKo-Bubbles,[15] visualisiert.

Schlussplädoyer

Korpora in eine eigene Studie miteinzubeziehen oder diese sogar darauf aufzubauen, löst sicher nicht alle Probleme – im Gegenteil: Es kommen stattdessen neue Herausforderungen hinzu. Man braucht geeignete Daten und eine Infrastruktur, um die Daten auszuwerten, und auch eine gewisse Offenheit für eine korpuslinguistische Denk- und Arbeitsweise. In dieser Übersicht wollten wir Ihnen einige Einstiegspunkte anbieten, neben den Recherchesystemen mit ihren Datenbeständen auch die Experimentierplattformen als Ausblick des Machbaren. Diese Angebote zu den eher allgemeinsprachlich gehaltenen Quer- und Ausschnitten der deutschen Sprache sind durchaus auch dann eine sinnvolle Ergänzung, wenn ein selbst erstelltes Korpus ausgewertet wird, wie es für diskursanalytische Fragestellungen naheliegend sein kann. Denn die These, dass sprachliche Phänomene (diskurs-)spezifisch ausgeprägt sind, benötigt zur Bestätigung die Erkenntnis, dass es bei diesen Phänomenen allgemeinsprachlich anders aussieht.

15 Einen guten Einstieg stellt die Dokumentation der letzten Studie (http://corpora.ids-mannheim.de/openlab/sliceviz/description.html (28. 1. 2022)) dar.

Literaturverzeichnis

Bodmer Mory, Franck (2014): Mit COSMAS II »in den Weiten der IDS-Korpora unterwegs«. In: Institut für Deutsche Sprache (Hrsg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Redaktion: Steinle, Melanie/Berens, Franz Josef. Mannheim: Institut für Deutsche Sprache. S. 376–385.

Kupietz, Marc/Schmidt, Thomas (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In: Eichinger, Ludwig M. (Hrsg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. Berlin/Boston: De Gruyter. S. 297–322.

Teubert, Wolfgang/Belica, Cyril (2014): Von der linguistischen Datenverarbeitung am IDS zur »Mannheimer Schule der Korpuslinguistik«. In: Institut für Deutsche Sprache (Hrsg.): Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Redaktion: Steinle, Melanie/Berens, Franz Josef. Mannheim: Institut für Deutsche Sprache. S. 298–319.

Rainer Perkuhn, Leibniz-Institut für Deutsche Sprache, Mannheim,
perkuhn@ids-mannheim.de