Originalveröffentlichung in: Deppermann, Arnulf/Fandrych, Christian/Kupietz, Marc/Schmidt, Thomas (Hrsg.): Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial. – Berlin/Boston: de Gruyter, 2023. S. 1-28. (Jahrbuch / Leibniz-Institut für Deutsche Sprache 2022)
DOI: *https://doi.org/10.1515/9783111085708*

Marc Kupietz/Harald Lüngen/Nils Diewald (Mannheim)

Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo

Vom Design bis zur Verwendung und darüber hinaus

Abstract: Das Deutsche Referenzkorpus DeReKo dient als eine empirische Grundlage für die germanistische Linguistik. In diesem Beitrag geben wir einen Überblick über Grundlagen und Neuigkeiten zu DeReKo und seine Verwendungsmöglichkeiten sowie einen Einblick in seine strategische Gesamtkonzeption, die zum Ziel hat, DeReKo trotz begrenzter Ressourcen für einerseits möglichst viele und andererseits auch für innovative und anspruchsvolle Anwendungen nutzbar zu machen. Insbesondere erläutern wir dabei Strategien zur Aufbereitung sehr großer Korpora mit notwendigerweise heuristischen Verfahren und Herausforderungen, die sich auf dem Weg zur linguistischen Erschließung solcher Korpora stellen.

1 Einleitung

Dieser Beitrag gibt einen Überblick über die Gesamtkonzeption des Deutschen Referenzkorpus DeReKo – von seinen Designprinzipien, über Ausbau- und Aufbereitungsstrategien, bis hin zur Erweiterung seiner linguistischen Nutzungsmöglichkeiten. Besonderes Augenmerk gilt dabei aktuellen Herausforderungen und der Vorstellung unserer Lösungsansätze, die jeweils durch eine enge Integration allgemein methodischer, linguistischer, informatischer und infrastruktureller Aspekte charakterisiert sind.

Im folgenden Abschnitt 2 werden kurz DeReKo's Aufgaben und Ziele, Designprinzipien und Erweiterungsstrategien zusammengefasst. Abschnitt 3 berichtet über die aktuelle Vorgehensweise bei der Akquisition und Aufbereitung von Texten und will außerdem auf einen in der Literatur bisher wenig explizit diskutierten Umstand aufmerksam machen: Die Forschungsdatenaufbereitung für sehr große Korpora wie DeReKo erfordert im großen Maßstab den Einsatz heuristischer Verfahren, was u. a. auch erhebliche Konsequenzen für die Methodik der Korpusnutzung hat. Dazu werden einige Beispiele dargestellt und die im Kontext von DeReKo angewendeten Lösungsstrategien skizziert. Abschnitt 4 berichtet über die jüngsten Ergebnisse der zuvor dargestellten

https://doi.org/10.1515/9783111085708-002

Ansätze: aktuelle DeReKo-Erweiterungen und Verbesserungen in der Abdeckung in den Bereichen Internetbasierte Kommunikation und Fachsprache. Im Abschnitt 5 geht es um die sich anschließende Herausforderung, wie trotz rechtlicher, methodischer, technischer und ökonomischer Grenzen sehr große Korpora wie DeReKo, für einerseits möglichst viele, andererseits aber auch für innovative und anspruchsvolle linguistische Anwendungen möglichst niedrigschwellig nutzbar gemacht werden können. Wir stellen dazu eine aktualisierte und verfeinerte Fassung unseres „put the computation near the data"-Ansatzes (Gray 2003; Kupietz et al. 2010) vor und gehen auf konkrete Verbesserung der Möglichkeiten programmatischer Nutzung ein, insbesondere für kontrastive und vergleichende Forschung.

2 DeReKo-Grundlagen

2.1 Aufgaben und Ziele

Das Deutsche Referenzkorpus DeReKo wird am Leibniz-Institut für Deutsche Sprache bereits seit dessen Gründung 1964 aufgebaut. Aufgabe und Ziel von DeReKo ist es, eine allgemeine Forschungsdatengrundlage für das IDS und für die synchron arbeitende germanistische Linguistik insgesamt dauerhaft zu sichern und dabei möglichst breit einsetzbar zu sein, z. B. für Forschung in den Bereichen Lexikographie, Grammatik und Orthographie über DaF, Forensische Linguistik, Diskurslinguistik bis zu Sprachkritik: Linguist/-innen und, sofern möglich, auch Forschende aus angrenzenden Disziplinen sollen durch DeReKo in die Lage versetzt werden, sich für eine große Bandbreite an Fragestellungen und Sprachdomänen geeignet stratifizierte Sub-Korpora zu definieren, mithilfe derer sie bestehende Hypothesen zuverlässig testen und interessante neue Hypothesen gewinnen können. Zu diesem Zweck wird DeReKo laufend stichprobenartig um ein möglichst breit gefächertes Spektrum des aktuellen deutschen Schriftsprachgebrauchs erweitert und mehrfach morphosyntaktisch und syntaktisch annotiert. Zuständig für DeReKo ist seit 2004 das IDS-Dauerprojekt *Ausbau und Pflege der Korpora geschriebener Gegenwartssprache*.

2.2 Urstichproben-Design: Stratifizierte nutzerdefinierte Korpora

Seit der Einführung von COSMAS I (al Wadi 1994) ist DeReKo einem *Urstichproben-Design* (Kupietz et al. 2010) verpflichtet, d. h. DeReKo gilt als eine Urstichprobe (engl. *primordial sample*) der deutschen Schriftsprache. DeReKo zielt somit in der Akquisitionsphase nicht auf eine formale Ausgewogenheit, wie es vielleicht von anderen Referenzkorpora bekannt ist, die nach einem bestimmten Schlüssel feste Anteile an Genres vereinen, wie das wegweisende British National Corpus (BNC Consortium 2007). Vielmehr strebt DeReKo eine möglichst breite Streuung und Besetzung potenziell relevanter Strata wie Zeit, Ort, Genre oder Thema an, um seine Nutzer in die Lage zu versetzen, sich aus DeReKo anhand seiner Metadaten selbst gezielt stratifiziert *virtuelle Korpora* zusammenzustellen, die bezüglich ihrer konkreten Forschungsfrage und Sprachdomäne eine geeignete und im besten Fall repräsentative Stichprobe darstellen.

2.3 Steuerung des DeReKo-Ausbaus

Bei der Steuerung des Ausbaus von DeReKo werden verschiedene Faktoren berücksichtigt, die wie bei einem Optimierungsproblem koordiniert werden müssen.

1. Die **Steigerung der Größe und Diversität** sind grundsätzliche Ziele, um den Status von DeReKo als Urstichprobe der schriftlichen Gegenwartssprache fortlaufend zu konsolidieren.
2. Insbesondere ist dabei auch die **Kontinuität** und **Aktualität** hervorzuheben, um (zeitnah) Sprachwandelprozesse erfassen zu können.[1]
3. Zur Gewährleistung der Kontinuität ist die **Wahrung des Renommees** des IDS als verlässlicher Partner für Text- und Lizenzspender notwendig.
4. Außerdem spielen **langfristige Strategien und Prognosen** (z. B. über die Ubiquität von Digitalisierung oder die Entwicklung der Presselandschaft) eine Rolle.
5. Besonders bzgl. der Diversitätsverbesserung wird auf die **Nachfrage** und den Bedarf von IDS-internen und gegebenenfalls externen Forschungsprojekten eingegangen.
6. Die Akquisition ist grundsätzlich abhängig vom tatsächlichen **Angebot** – es kann nur akquiriert werden, was auf der Seite von Textgebern und Rechte-

1 Siehe auch Abschnitt 3 zum entsprechenden Satzungsauftrag des IDS.

inhabern (wie Zeitungs- und Buchverlagen, Datenbankprovidern, Portalbetreibern) sowie Forschungseinrichtungen oder Einzelpersonen, die selbst Korpora aufbauen, angeboten wird.

7. Die Datenakquisition wird auch priorisiert anhand der anfallenden **Kosten** für Verhandlungsaufwand und Lizenzgebühren sowie für die anschließende Erschließung (Aufwand an Analyse, Konvertierung und Aufbereitung zur Integration in DeReKo) und Wartung.

DeReKo wird zwei Mal im Jahr aktualisiert und in Form eines sogenannten DeReKo-Releases veröffentlicht, das daraufhin in die Korpusrecherchesysteme COSMAS II (Bodmer 1996; b. a. w.) und KorAP eingepflegt wird.

3 Herausforderungen der Forschungsdatengewinnung

Viele Herausforderungen, die sich bei der Erweiterung von DeReKo ergeben, sind unmittelbar auf seine Größe und sein Wachstum zurückzuführen. Der Stichprobenumfang ist jedoch ein entscheidender Faktor für die Verallgemeinerbarkeit ihrer Eigenschaften und für Gewinnung interessanter linguistischer Erkenntnisse. „More data are better data" (cet. par.) gilt in der Linguistik mehr noch als in vielen anderen Disziplinen, da lexikalische Häufigkeitsverteilungen eine *large number of rare events* (*LNRE*) aufweisen, mit linguistisch interessanten Phänomenen oft weit hinten im sogenannten *long tail* (vgl. Kupietz/Schmidt 2015, S. 302). Hinzu kommt, dass sprachliche Variation von vielen inner- und außersprachlichen Kontextvariablen abhängt, so dass auch in sehr großen Korpora Beobachtungen zu bestimmten relevanten Kombinationen dieser Variablen rar sein können.

Unabhängig von solchen methodischen Überlegungen leitet sich die Notwendigkeit der kontinuierlichen DeReKo-Erweiterung, speziell um aktuelle Daten, auch aus dem Stiftungszweck des IDS ab: „Die Stiftung verfolgt den Zweck, die deutsche Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte wissenschaftlich zu erforschen und zu dokumentieren." (Leibniz-Institut für Deutsche Sprache 2020, § 2(1)).

3.1 Korpusakquisition

Der typische Workflow einer Akquisitionskampagne zur Erweiterung von DeReKo beginnt mit der Identifikation eines Texttyps oder Stratums entsprechend der

oben genannten Kriterien, für das Texte neu akquiriert werden sollen (wie beispielsweise Belletristik). Sodann werden 50–100 potenzielle Textgeber ermittelt, und es wird versucht passende Ansprechpartner (z. B. die Leitung der Öffentlichkeitsarbeit oder Lizenzabteilung eines Verlags) herauszufinden. Erfolgversprechend ermittelte Personen erhalten per Post ein Anschreiben durch den Wissenschaftlichen Direktor des IDS mit einführenden Informationen über DeReKo und einem Antwortformular. Die Erfahrung zeigt, dass sich darauf ca. 5% der Angeschriebenen mit einer positiven Antwort zurückmelden. Bei diesen kann danach, vorwiegend telefonisch, genauer geklärt werden, welche Texte und wieviele in welchen Formaten zu welchen Lizenzbedingungen zur Verfügung gestellt werden können. Zumeist kann im Anschluss anhand dieser Angaben und idealerweise anhand der Sichtung von Beispieldaten eine gute Kosten-Nutzen-Abschätzung der Quelle durchgeführt werden, d. h. Einschätzungen darüber, wie aufwändig die Konvertierung in das DeReKo-Datenformat I5 sein wird, was die Quelle langfristig kosten wird, welchen linguistischen Nutzen sie bringt (z. B. bzgl. der Erschließung neuer Strata oder durch möglicherweise interessante ggf. rekonstruierbare Metadaten). Bei den Lizenzverhandlungen geht es primär um Faktoren wie die Kosten und Laufzeit der Lizenz, die Höhe der Aufwandsentschädigung, sowie darum, ob die Lizenz auch auf mögliche zukünftige Datenlieferungen übertragbar sein soll. In der Vergangenheit haben viele Textgeber die unveränderte DeReKo-Standard-Lizenzvereinbarung abgeschlossen, in letzter Zeit gab es häufig noch besondere Wünsche seitens der Rechtsabteilungen der Verlage bzgl. des Wortlauts der Vereinbarungen.

3.2 Korpusaufbereitung

Neu akquirierte Korpora, insbesondere solche von bisher nicht zu DeReKo beitragenden Datengebern, werden in der Regel in Formaten übermittelt, die zwar in XML vorliegen, jedoch zunächst Anpassungen bedürfen, um sie in DeReKo aufnehmen zu können. Hierfür wird als erstes ein Abgleich mit bestehenden Datenformaten anderer Datengeber unternommen und als Basis der Anpassungen jene Konvertierungsroutinen ausgewählt, die dem Eingangsformat am ehesten entsprechen. Damit für unterschiedliche Rohdatenformate nicht jeweils vollständig separate Konvertierungsroutinen entwickelt und gewartet werden müssen, wurde ein hierarchisches Konvertierungsmodell auf Basis kleiner und wartbarer XSLT-Skripts entwickelt (Kupietz/Keibel 2009), in dem die Funktionalitäten der übergeordneten allgemeineren Ebenen (z. B. generell für Presseerzeugnisse) auf den darunterliegenden Ebenen (z. B. Redaktionssystem, Verlag, spezielle Zeitung) jeweils geerbt, ggf. überschrieben und verfeinert werden können.

Bei der Konvertierung der Daten wird das Hauptaugenmerk auf die Überführung und eventuelle Rekonstruktion geeigneter Metadaten gelegt, die in das Metadatenmodell von DeReKo passen und für die Korpuskomposition und für die Interpretation der Phänomen-/Trefferverteilungen von Relevanz sind. Eine Angleichung an das Textsortenmodell stellt hierbei eine besondere Herausforderung dar. Die Interpretation dieser Eingangsmetadaten erfordert im Allgemeinen eine eingehende Datensichtung und lässt sich nur bedingt automatisieren. Dies gilt auch für die Überführung der Textstrukturinformationen (siehe Abschn. 3.3).

Neue Korpora und ihre Bestandteile müssen zudem auf das hierarchische IDS-Textmodell[2] abgebildet werden, bei dem eine Entscheidung getroffen werden muss, welche Zuordnung auf Korpus-, Dokument- und Textebene erfolgen soll.

Bei Korpora, die kontinuierlich erweitert werden, wie dies bei Zeitungen und Zeitschriften der Fall ist, müssen die Konvertierungsroutinen regelmäßig neu evaluiert werden (üblicherweise vor einem DeReKo-Release), da Datengeber ihre Formate ständig ihren Ansprüchen folgend erweitern und anpassen. Im besten Fall führt dies zur Einführung neuer Metadaten oder neuer Textsorten, die mit Informationsgewinn in das Metadatenmodell von DeReKo eingefügt werden können. Im schlechteren Fall ist die etablierte Konvertierungsroutine inkompatibel und muss neu aufgebaut werden.

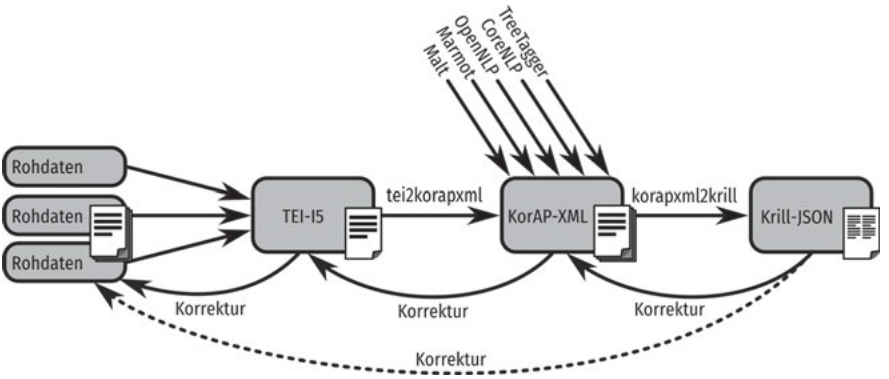

Abb. 1: Aufbereitungspipeline für DeReKo

2 https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell/ (Stand: 3.8.2022).

Infolge etwaiger Anpassungen der Konvertierungsskripte müssen unter Umständen auch in den weiteren Aufbereitungsschritten (siehe Abb. 1) Anpassungen vorgenommen werden, z. B. um neue Metadatenkategorien bzw. Variablen zu erfassen, die mitunter neue linguistische Anwendungsgebiete eröffnen. Sobald alle Aktualisierungen vorgenommen wurden, kann die Aufbereitung für ein neues DeReKo-Release gestartet werden. Diese läuft in mehreren Schritten und parallel auf mehreren Rechnern ab.

Zunächst werden die Rohdaten der einzelnen Datengeber mittels der angepassten Konvertierungsroutinen in das einheitliche Format I5 konvertiert (Lüngen/Sperberg-McQueen 2012). I5 ist das primäre Repräsentationsformat von DeReKo und dient inzwischen für viele interne Analysewerkzeuge als Eingangsformat, unter anderem für COSMAS II. Allerdings erlaubt dieses Format Annotationen nur in begrenztem Maße, so werden beispielsweise konkurrierende Annotationen nicht unterstützt.

Um diese Beschränkung aufzubrechen, werden in einem weiteren Schritt mit dem Werkzeug tei2korapxml (Harders et al. 2020–) die Daten in das interne KorAP-XML-Format (Bański et al. 2012) überführt. Je nach Eingangskorpus sind hier unterschiedliche Konfigurationen vorzunehmen, da insbesondere in anderen Projekten entwickelte Korpora gelegentlich vorannotiert sind oder vom DeReKo-Metadatenmodell abweichen und entsprechend gesondert konvertiert werden müssen. Diese Korpora sind zumeist allerdings statisch und bedürfen keiner kontinuierlichen Anpassung. Das KorAP-XML-Format erlaubt es, beliebige Annotationen den Primär- und Metadaten unabhängig (d. h. auch zeitgleich in paralleler Verarbeitung) hinzuzufügen. In diesem Schritt werden die Daten auch mit dem KorAP-Tokenizer (Kupietz/Diewald 2022; Diewald/Kupietz/Lüngen 2022) tokenisiert, sofern noch keine Tokenisierung vorhanden ist. In der Regel ist die Textstruktur die einzige Annotationsebene, die direkt aus den I5-Quellen übernommen wird.

In Bezug auf Rechenzeit- und Ressourcen-Bedarf ist die automatisierte Annotation der aufwändigste Schritt der Korpusaufbereitung (vgl. Belica et al. 2011). Für die Wortartenerkennung werden derzeit OpenNLP,[3] TreeTagger (Schmid 1994), CoreNLP (Manning et al. 2014) und MarMoT (Müller/Schmid/Schütze 2013) eingesetzt, für die Lemmatisierung TreeTagger, für morphologische Annotationen MarMoT, für komplementäre Satzgrenzenerkennung OpenNLP und CoreNLP, für Konstituenzannotationen CoreNLP und für Dependenzannotationen MaltParser (Hall/Nivre 2008). Alle Annotationen werden in KorAP-XML zusammengefasst.

3 https://opennlp.apache.org/ (Stand: 3.8.2022).

Nach der Anreicherung aller Daten ist für die Indizierung in KorAP ein weiterer Konvertierungsschritt mit dem Werkzeug korapxml2krill (Diewald 2016–) notwendig, in dem die separiert vorliegenden Annotationen zusammengefasst werden und pro Text eine hochannotierte Datei erzeugt wird.

In allen Zwischenschritten werden Log-Informationen hinsichtlich Verarbeitungsfehler kontrolliert und gegebenenfalls Korrekturen in den Verarbeitungsskripten vorgenommen, da viele Probleme und Inkompatibilitäten der Eingabedaten erst im Zuge der Verarbeitung auftreten. Um (partielle) Duplikate in den Daten zu kennzeichnen, wird zusätzlich eine Dubletten-Detektion durchgeführt (Kupietz 2005). Bei Aufbereitungsfehlern, die schon in vergangenen DeReKo-Releases auftraten, ist abzuwägen, ob diese für neue Veröffentlichungen korrigiert (und damit Unstetigkeiten in den Daten über Releases hinweg einführen) oder lediglich dokumentiert werden. Es ist auch möglich, dass ein Fehler, der in einem frühen Schritt der Verarbeitung entsteht, erst spät auffällt, was gegebenenfalls eine vollständige Neuverarbeitung erforderlich macht (siehe Korrekturpfeile in Abb. 1). Diese Trial-and-Error-Zyklen können bei großen Korpora sehr lang sein, weshalb Rechenzeit bei der Korpusaufbereitung trotz aller Automatisierung eine nicht zu vernachlässigende Größe ist und zu erheblichen Verzögerungen bei DeReKo-Releases führen kann.[4]

Nicht immer sind Probleme auf die Korpusdaten und entsprechend auf die Konvertierungsroutinen zurückzuführen. Herausforderungen stellen auch technische Hürden während der Aufbereitung dar, etwa hinsichtlich des benötigten Arbeitsspeichers, des Festplattenplatzes, der Limitierung von Einträgen in Dateisystemen oder Netzwerkausfälle bei stark parallelisierten Verfahren. Ohne Verteilung auf mehrere Rechner und mehrere Prozessoren würde die Aufbereitung eines DeReKo-Releases derzeit etwa zwei Jahre Rechenzeit benötigen.[5] Entsprechend anspruchsvoll ist auch die Ausführung und Koordinierung der Abläufe, die nicht vollständig automatisiert und ohne (Nach-)Kontrolle ablaufen kann.

Nach der Vorverarbeitung können die Daten in COSMAS II, KorAP und weiteren Analysewerkzeugen indiziert werden. Auch für diesen Schritt gelten oben genannte Herausforderungen. Dynamische virtuelle Korpora, denen neue Korpora zugehörig sind, werden dabei aktualisiert.

---

4 Die Konvertierung der Wikipedia-Korpora kann beispielsweise bis zu einer Woche dauern. Ein allein in diesem Korpus erst spät aufgefallener Fehler hat in der Vergangenheit bereits zur deutlichen Verschiebung einer Veröffentlichung von DeReKo geführt.

5 Zurzeit werden 14 Unix-Rechner mit insgesamt über 200 CPU-Cores eingesetzt.

3.3 Notwendigkeit heuristischer Verfahren und resultierende Herausforderungen

Die Vorteile der Nutzung sehr großer Datenmengen wird oft mit dem Begriff „Big Data“ in Verbindung gebracht, den Laney (2001) als das Wachstum von Daten (-beständen) in den drei Dimensionen *volume*, *velocity* und *variety* definiert (3V-Modell). In diesem Sinne kann auch DeReKo als Big Data aufgefasst werden, denn sowohl Größe, Wachstum als auch Stratifizierung sind Teil der Kernstrategie seines Aufbaus. Die damit einhergehenden Nachteile hinsichtlich der Datenaufbereitung und Datenauswertung, insbesondere die Notwendigkeit speziell darauf ausgelegter Software und Methoden (vgl. Liu et al. 2016), haben zur Folge, dass oft nicht die qualitativ beste, sondern lediglich die praktikabelste Lösung für eine Aufgabe eingesetzt werden kann. So werden für die automatische Annotation nur jene Werkzeuge eingesetzt, die DeReKo in akzeptabler Zeit und mit den im Projekt verfügbaren Ressourcen verarbeiten können. Zudem können für die Analyse von Rohdaten oder die etwaiger Verarbeitungsfehler nur Stichproben und heuristische Verfahren eingesetzt werden, da der Datenumfang und seine kontinuierliche Erweiterung eine exakte Durchsicht unmöglich macht.[6]

Die Notwendigkeit des Einsatzes von Heuristiken macht dabei einen qualitativen Unterschied für die Korpusaufbereitung aus – im Vergleich zur Aufbereitung anderer Sammlungen objektiv messbarer Daten. Sobald heuristische Verfahren verwendet werden, ist Wissen über das Korpus und sein linguistisches Anwendungsspektrum notwendig, da aufbereitete Korpora nicht mehr anhand ihrer Korrektheit bewertet werden können (was korrekt ist, ist unbekannt), sondern anhand des Verhältnisses ihrer Tauglichkeit für die intendierten Anwendungen zum investierten Aufwand. Das heißt, dass in diesem Fall nicht nur anspruchsvollere informatische Entscheidungen getroffen werden, sondern auch allgemeine methodische, zum Beispiel in Hinblick auf die Homogenität der Daten, und speziell linguistische, die sich meist nicht in richtig oder falsch kategorisieren lassen.

Ein sehr grundlegendes Beispiel, das wenig Beachtung findet, aber weitreichende Konsequenzen für fast alle linguistischen Untersuchungen hat, ist die Tokenisierung und Satzsegmentierung (siehe auch Diewald/Kupietz/Lüngen 2022; Diewald 2022). Nicht allgemein zu beantworten, aber trotzdem recht all-

6 DeReKo wurde 2020 pro Arbeitstag durchschnittlich um den Umfang von über 100 Spiegel-Ausgaben oder 35-mal den Roman „Buddenbrooks“ erweitert. Für eine exakte Durchsicht und Verarbeitung dieses Umfangs müsste die Anzahl der Projektmitarbeiter/-innen etwa verhundertfacht werden.

gemein zu entscheiden ist dabei z. B., ob ein Punkt eine Abkürzung oder ein Satzende markiert, ob eine Zeichensequenz ein Emoticon darstellt, wann ein Asterisk als Gendersternchen gemeint ist und wie Mehrwortausdrücke zu behandeln sind.

Zusätzlich zu solchen meist auf einer generellen Ebene, wie z. B. für das gesamte Korpus oder vielleicht abhängig von Medialität oder Genre, zu klärenden Punkte, gibt es auch viele Fragen, die auf einer spezifischeren Ebene, wie z. B. bzgl. einer bestimmten Textquelle, zu beantworten sind. Ein typischer Anwendungsfall ist dabei die tentative Dekodierung von ambigen visuell-optischen Auszeichnungen (vgl. Perkuhn/Keibel/Kupietz 2012, S. 55) zur Rekonstruktion struktureller Eigenschaften von Textpassagen, wie z. B. zur Segmentierung und Auszeichnung von Überschriften, anhand von Stilattributen wie Schriftstärke, -größe und Zeilenvorschüben. Auswirkungen auf linguistische Anwendungen haben solche Heuristiken nicht nur, wenn ausschließlich in Überschriften gesucht oder diese explizit ausgeschlossen werden sollen, sondern auch, wenn nur die Segmentierung etwa bei der Untersuchung von Mehrwortausdrücken eine Rolle spielt.

Fehler und generell unerwartetes oder inkohärentes Verhalten sind darüber hinaus in aggregierten Darstellungen wie Kookkurrenzanalysen kaum noch ermittelbar. Zudem gilt natürlich generell, dass falsch Negative naturgemäß nicht erkennbar sind.[7]

Während die Relevanz der Problematik der Dekodierung von visuellem Markup durch die zunehmende Verwendung von generischem Markup zumindest in Rohdaten, die aus Redaktionssystemen von Tageszeitungen stammen, perspektivisch abnimmt, stellt die heuristische Ermittlung von extratextuellen Variablen, bzw. Metadaten, zu einzelnen Texten eine allgegenwärtige Herausforderung dar. Typische Beispiele sind die automatische Zuordnung und Vereinheitlichung von Textsorten und Zeitungsressorts und die thematische Klassifikation von Texten. Die spezifische Herausforderung bei der vereinheitlichten Kategorisierung von Ressorts und Zeitungsartikeltypen (z. B. Agenturmeldung vs. Kommentar) ist, dass die zugrundeliegenden Originalmetadaten (die meist zusätzlich ausgezeichnet werden) sehr volatil sind. Bei der Entwicklung diesbezüglicher Heuristiken sind also auch Aspekte der Wartbarkeit bzw. der Homogenität der (Meta-)Daten in Hinblick auf die zukünftige Konstruierbarkeit virtueller Korpora und multidimensionale Analysen zu berücksichtigen. Fehler und entsprechende Schwankungen in den Daten sind jedoch im Fall von DeReKo unvermeidbar.

7 Siehe Belica et al. (2011) für eine detaillierte Diskussion der Problematik von Fehlern zweiter Art.

Einen besonderen Fall stellt DeReKo's thematische Textklassifikation dar, die sich im Kontinuum zwischen Beobachtungsaufzeichnungen und Interpretationen weit auf der Seite der Interpretationen befindet. Es werden dazu keine gegebenen Metadaten herangezogen. Die Klassifikation eines Textes erfolgt anhand seines Vokabulars durch einen auf annotierten Daten trainierten automatischen Klassifikator bzgl. einer zweistufigen Teilmenge der Open-Directory-Taxonomie (dmoz, siehe Klosa et al. 2012). Ähnlich wie bei anderen kategorialen Variablen, z. B. Textsorte und Genre, ist das zugrundeliegende Kategoriensystem im Prinzip arbiträr. Standards dazu existieren mit dem Dewey Decimal Classification System (DDC) und der Universellen Dezimalklassifikation (UDC) vor allem im Bibliotheksbereich. Diese sind jedoch für die Einteilung von Wissensgebieten konzipiert, so dass sie große Teile von DeReKo nicht abdecken. Diesbezüglich besser geeignet und entsprechend vielversprechender auch im Hinblick auf eine Etablierung als De-Facto-Standard in der Korpuslinguistik, erscheinen die thematischen Top-Level-Kategorien der Wikipedia, die z. B. vom Referenzkorpus der Rumänischen Gegenwartssprache CoRoLa (Tufiș et al. 2016), neben UDC, verwendet werden (Gifu et al. 2019).

3.4 Lösungsstrategien

Mit Fehlern rechnen und umgehen

Große und hinsichtlich vieler Variablen breit gestreute Korpora sind für eine detaillierte Erforschung des Sprachgebrauchs unerlässlich. Die dazu benötigte Forschungsdatengewinnung und Aufbereitung ist auf die Verwendung von Heuristiken angewiesen. Die damit verbundenen Fehler sind nicht vollständig vermeidbar. Der im Kontext von DeReKo seit langem propagierte Lösungsansatz besteht daher vor allem darin, auf Fehler vorbereitet zu sein und mit diesen möglichst gut umzugehen. Auf der Seite der Korpusnutzung heißt das allgemein, dass erste Schlussfolgerungen aus Korpusbefunden als Hypothesen zu betrachten sind, was aber auch sonst bei kleinen, sorgfältig manuell erstellten Korpora ratsam ist.

Ein sinnvoller Umgang mit den erwarteten Fehlern bedeutet etwa bei der Konstruktion von virtuellen Korpora, iterativ Samplingfehler zu korrigieren (Kupietz 2015) oder Suchanfragen iterativ so anzupassen, dass falsch negative Treffer ausgeschlossen und falsch positive Treffer minimiert sind (Belica et al. 2011) – jeweils unabhängig davon, ob die zunächst beobachteten Fehler auf eine fehlerhafte Datenaufbereitung zurückzuführen sind oder nicht.

Konzentration auf linguistisch relevante Fehler

Auf der Seite der Korpusaufbereitung besteht, wenn das Ziel einer vollständigen Fehlervermeidung ohnehin nicht erreichbar ist, meist ein viel unmittelbarerer Tradeoff-Effekt des investierten Aufwands auf die erreichbare Korrektheit. Es lohnt daher, sich bei der Korpusaufbereitung auf die Vermeidung solcher Fehler zu konzentrieren, die für viele linguistische Anwendungen relevant sind – was allerdings die Kenntnis dieser voraussetzt. Analog kann in der Anwendung ein virtuelles Korpus möglicherweise so eingeschränkt werden, dass ein für diese Anwendung relevanter Fehler umgangen wird, sofern dadurch die Stichprobe nicht verzerrt wird, was wiederum eine Kenntnis des Korpus und das Wissen um eingesetzte Heuristiken erfordert.

Im Zweifel weitere Meinungen einholen

Ein weiterer genereller Ansatz zum Umgang mit Fehlern und Unsicherheiten besteht darin, im Zweifel sozusagen mehrere Meinungen etwa durch die Verwendung unterschiedlicher Tools beispielsweise bei der Klassifikation von Wortarten (Belica et al. 2011) oder bei der Zuordnung von Themen zu Texten einzuholen, um anhand der Abweichungen unter den Ratings der Klassifikationstools einen Überblick über potenzielle Problembereiche zu erhalten und ggf. wahlweise Präzision oder Recall zu maximieren (vgl. Kupietz et al. 2017).

Anwendungsspezifische ad-hoc Metadaten und Annotationen

Dieser obige Ansatz ähnelt einem weiteren, der häufig zur Anwendung kommt. Er besteht darin, die Qualität von bestimmten Metadaten oder Annotationen anlässlich eines bestimmten Anwendungsfalls zu verbessern oder neue Metadatenkategorien für diesen hinzuzufügen. Häufig wird dieser Ansatz verwendet, wenn für ein bestimmtes Projekt etwa ein abweichendes Kategoriensystem von Textsorten benötigt wird oder das bestehende für ein bestimmtes Subkorpus genauer sein muss. Möchte ein Projekt z. B. die Veränderung des Sprachgebrauchs speziell in Zeitungsinterviews über die Zeit beobachten, kann es zur Erhöhung des Recalls für virtuelle Korpuskonstruktion zusätzlich zum Metadatum für den Artikeltyp Interview weitere Kriterien heranziehen. Für den Spiegel, z. B., könnte eine Heuristik so aussehen, dass eine bestimmte Häufigkeit von Sätzen, die mit „SPIEGEL:“ beginnen, verlangt wird. Gerade bei virtuellen Korpora, die sich über größere Zeiträume erstrecken, lässt sich so die Qualität deutlich verbessern, bzw. im

Beispiel das virtuelle Korpus deutlich vergrößern. Das Beispiel macht jedoch auch eine ganze Reihe typischer Anschlussfragen deutlich: Soll die projektspezifische Heuristik in die allgemeinen Aufbereitungswerkzeuge integriert werden? Ist die Heuristik und ihre Implementation wartbar? Welche negativen Effekte hätte sie für andere Anwendungen? Sind über alle Verwendungen betrachtet die ursprünglichen Werte des Artikeltyp-Metadatums besser geeignet? Falls ja, lohnt es sich, eine neue Metadatenkategorie einzuführen? Wäre diese ausreichend allgemein, damit auch andere Anwendungen davon profitieren können? Wäre der nötige Aufwand gerechtfertigt und zu bewältigen? Oder sollte der Artikeltyp stattdessen mehrere Zuordnungen zulassen? Falls ja, können alle verwendeten Analysewerkzeuge mit einer solchen Änderung umgehen? Bei allen Änderungen: Können diese auch auf bereits veröffentlichte Daten angewendet werden, ohne die Reproduzierbarkeit von Forschungsergebnissen zu gefährden? Können die Änderungen nur für zukünftig zu veröffentlichende Daten gemacht werden, ohne eine potenziell irreführende Unstetigkeit in den Daten einzuführen?

Da die wissenschaftlichen (und ökonomischen) Folgen oft weitreichend und schwer überschaubar sind, werden solche aus speziellen Projekten hervorgehenden Anreicherungen oder auch potenziellen Verbesserungen von DeReKo häufig nicht auf veröffentlichte Daten angewendet oder für zukünftige Konvertierungen verwendet. Stattdessen werden die Anreicherungen getrennt vom Korpus (standoff) vom jeweiligen Projekt gespeichert, wobei die eindeutige Referenz zu Texten über Text-IDs (Siglen) hergestellt wird. Idealerweise ist zusätzlich zu dieser statischen, extensionalen Variante noch eine Operationalisierung verfügbar, die eine Anwendung auf zukünftige Daten im Prinzip möglich macht. Der Nachteil dieser Vorgehensweise ist, dass andere DeReKo-Nutzer/-innen, von solchen Anreicherungen nicht ohne Weiteres profitieren können und Operationalisierungen nicht in den Wartungsprozess der DeReKo-Aufbereitung einbezogen werden.

Heuristiken zur aktiven Detektion von Fehlern

Heuristiken zur aktiven Erkennung von Fehlern sind besonders dann eine ökonomisch sinnvolle Ergänzung zu spezifischen Aufbereitungsheuristiken, wenn sie möglichst allgemein, also möglichst für das gesamte Korpus einsetzbar sind. Ansatzpunkt für solche Heuristiken sind daher insbesondere quantitative Eigenschaften von Subkorpora, bzw. der Vergleich dieser Eigenschaften von Korpusneuzugängen mit Referenzwerten. Für DeReKo werden solche Techniken nur noch zur Detektion der häufigsten Fehlerklassen eingesetzt: Fehler in den Originaldatenlieferungen, fehlende Leerzeichen, falsche Zeichen-Enkodierungen und – in einem etwas anderen Kontext – zur Detektion und Auszeichnung von

Dubletten (siehe Abschn. 3.3). Gegen eine Ausweitung des Einsatzes solcher Techniken spricht, dass auch bei sehr allgemeinen Klassifikatoren es zu aufwändig wäre, die Heuristiken so einzustellen und zu warten, dass bei einer akzeptablen Zahl von falsch Negativen die Anzahl der falsch Positiven in einem zu bewältigenden Rahmen bleibt. Bei der automatischen Kontrolle der Datenlieferungen für DeReKo hat sich zum Beispiel gezeigt, dass die vom Überprüfungswerkzeug per E-Mail an die Projektmitarbeiter/-innen verschickten Fehlermeldungen nach kurzer Zeit wegen zu vieler falscher Alarme ignoriert werden und die eigentlich notwendige permanente Anpassung der Heuristik nicht realisierbar ist. Wiederum zeigt sich, dass bei ausreichend großen Datenmengen auch intuitiv als unwahrscheinlich eingeschätzte Probleme auftreten können.

Softwaretests zur Vermeidung von Fehlern

Von den verschiedenen Qualitätssicherungsmaßnahmen im Research Software Engineering (vgl. Diewald/Margaretha/Kupietz 2021) soll an dieser Stelle nur auf automatisierte Testtechniken eingegangen werden. Da sie sich nur in ihrer Implementation von den oben beschriebenen Heuristiken zur Fehlerdetektion unterscheiden und im Prinzip die gleiche Wartungsproblematik mit sich bringen, werden integrierte Softwaretests für die DeReKo-Aufbereitung ebenfalls nur für bestimmte Problemklassen eingesetzt. Eine davon betrifft die Kontrolle der Konvertierung als besonders problematisch identifizierter Rohtexte durch so genannte Regressionstests. Zur Kontrolle einer neu zu entwickelnden Aufbereitungsroutine wird z. B. nach und nach eine Stichprobe aus Rohtexten entwickelt und erweitert, die bei der Konvertierung Probleme verursacht haben. Die Kontrolle der Konvertierung dieser Stichprobe wird dabei in die Aufbereitungsroutinen so integriert, dass sie bei folgenden Releasezyklen nicht durch neue Tests für neue Daten ersetzt, sondern um diese erweitert und so immer wieder aufgerufen werden, so dass diesbezügliche Regressionen ausgeschlossen werden können.

Solche Regressionstests sind unmittelbar wichtig, da mögliche Probleme – angesichts von über 150 Pressequellen – auch dann nicht vollständig überschaubar sind, wenn sie zu einem früheren Zeitpunkt bekannt waren.

Ein weiterer Anwendungsfall für Regressionstests hängt mit der starken Hierarchisierung der Aufbereitungssoftware in viele, immer spezialisiertere Vererbungsebenen zusammen (siehe Abschn. 3.2). Der Vorteil dieses Ansatzes, nämlich dass für neue Rohdatenquellen oft nur wenig neuer Programmcode entwickelt werden muss, weil das meiste aus übergeordneten Klassen geerbt werden kann, steht dem kleineren Nachteil gegenüber, dass Änderungen auf

höheren Hierarchieebenen unerwünschte und unerwartete Konsequenzen haben können, die möglichst durch allgemeinere Regressionstests abgefangen werden sollten. Eine Erweiterung um mehr Tests dieser Art ist geplant.

Dokumentation bekannter Fehler

Bei Korpora der Größe von DeReKo ist notwendigerweise auch die Art der Dokumentation nicht vergleichbar mit der kleinerer, manuell aufbereiteter Korpora. Eine detaillierte Beschreibung aller Korpuseigenschaften wäre zu lang, als dass jemand diese schreiben oder lesen könnte. Ebenfalls gewöhnungsbedürftig im Korpuszusammenhang ist vielleicht schon die Idee, Fehler lediglich zu dokumentieren, statt sie vollständig zu vermeiden oder zu korrigieren. Im Fall von DeReKo können jedoch nicht alle Fehler sofort und manche auch nie korrigiert werden. Um bekannte Fehler zumindest nicht zu vergessen, sondern sie bei sich bietender Gelegenheit erneut zu sondieren und andere Nutzer/-innen auf diese hinzuweisen, werden DeReKo-Fehler seit einiger Zeit im Ticketsystem eines speziellen, internen Repositoriums des IDS-gitlab verwaltet. Dies bringt diverse Vorteile mit sich: Fehler können z. B. durchsucht, gelabelt, kommentiert mit Screenshots versehen, Meilensteinen und Mitarbeiter/-innen zugeordnet werden etc.[8]

4 Aktuelle DeReKo-Entwicklungen

4.1 Allgemeine Entwicklungen

Im Januar 2022 erschien das Release DeReKo-2022-I, welches nunmehr 52,97 Milliarden laufende Wörter enthält, wovon 96% (50,91 Mrd.) nach einer Registrierung öffentlich zugänglich sind. 2,06 Milliarden können aus lizenzrechtlichen Gründen nur intern an einem Arbeitsplatz im IDS verwendet werden.

Einen hohen Anteil am Wachstum haben die zahlreichen fortlaufend bereitgestellten Pressequellen. Dank der Zusammenarbeit mit vielen Einzelverlagen sowie seit 2013 mit einem großen deutschen Pressearchiv ist der deutschsprachige Raum durch diese Quellen mittlerweile reichlich und recht gleichmäßig

8 Allerdings ist das IDS-gitlab und damit auch das DeReKo-Ticketsystem leider aus datenschutzrechtlichen und organisatorischen Gründen derzeit nur IDS-intern zugänglich.

abgedeckt, d. h., die Diversität hinsichtlich der Dimension geografische Herkunft ist kontinuierlich hoch. In DeReKo-2018 kamen außerdem viele Publikums- und Fachzeitschriften hinzu und trugen zu einer höheren Diversität der Texttypen und Themengebiete bei. Seit 2018 wird dank der Kooperation mit einem Jugendbuchverlag das Korpus Kinder- und Jugendliteratur (Korpussigle kjl) angeboten und erhöht somit die Diversität der Zielgruppen in DeReKo.

4.2 Abdeckung Internetbasierter Kommunikation

Korpora Internetbasierter Kommunikation (IBK) spielen eine große Rolle für die Untersuchung u. a. von Neologismen und gesellschaftlichen Diskursen. DeReKo bemüht sich um eine fortlaufende Vergrößerung und Diversifizierung dieses Bereichs (Lüngen/Kupietz 2020). Hier sind zwei größere Neuerungen in DeReKo-2022-I zu verzeichnen: Zum einen wurde das NottDeuYTSch-Korpus (Nottinghamer Korpus deutscher YouTube-Kommentare; Korpussigle NDY) von Louis Cotgrove (2022, im Dr.) mit 33,7 Millionen Tokens in 3,1 Millionen YouTube-Kommentaren integriert.

Zum anderen kam das vom Korpusausbau-Projekt selbst erstellte Twitter-Sample-Korpus 2021 (Korpussigle TWI21) hinzu, welches 48 Millionen Wörter in 2,8 Millionen Tweets (eine Zufallsauswahl deutschsprachiger Tweets) ab dem 1.3.2021 enthält. Möglich wurde dies durch die Twitter API v2.0 und Twitters neue *Academic Research Track License*, die seit dem 26.1.2021 in Kraft ist (Kamocki et al. 2021).

Aufgrund der bestehenden rechtlichen Beschränkungen sind sowohl das NottDeuYTSch-Korpus wie auch das Twitter-Sample-Korpus bis auf Weiteres nur IDS-intern oder im Rahmen von Kooperationen nutzbar. Bisher hatte das Korpusausbau-Projekt aufgrund der Abwägung von Kosten für das Projekt und Nutzen für die wissenschaftliche Öffentlichkeit weitgehend darauf verzichtet, Korpora in DeReKo zu integrieren, die nicht entsprechend der üblichen Lizenzregelungen auch außerhalb des IDS abfragbar und analysierbar sind. Durch die am 7. Juni 2021 in Kraft getretene Novellierung der sogenannten Text-and-Data-Mining-Schranke (§ 60d UrhG) haben sich die Voraussetzungen jedoch dahingehend verbessert, dass solche Daten auch ohne Lizenz für eine uneigentliche, korpuslinguistische, dem Text-Mining entsprechende Nutzung „1. einem bestimmt abgegrenzten Kreis von Personen für deren gemeinsame wissenschaftliche Forschung sowie 2. einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung“ (§ 60d Abs. 4 UrhG) zugänglich gemacht werden dürfen und zudem die Regelungen zum Löschen der Daten liberalisiert wurden. Quantitative Auswertungen in diesem Sinne werden auch für die externe Nut-

zung, auch für die beiden o. g. Korpora, über die KorAP-API ermöglicht (siehe Abschn. 5.3).[9]

4.3 Weitere Erweiterungen

Ein weiterer Neuzugang ist das Korpus Gingko (*Geschriebenes Ingenieurwissenschaftliches Korpus*) des Projekts *Muster in der Sprache der Ingenieurwissenschaften* der Universitäten Greifswald und Leipzig. Es enthält die Jahrgänge 2007–2016 der Fachzeitschriften Automobiltechnische Zeitschrift (Korpussiglenpräfix ATZ) und Motortechnische Zeitschrift (Korpussiglenpräfix MTZ) des Springer Fachmedien-Verlags und umfasst 4,67 Millionen Tokens (Schirrmeister et al. 2021). Dieses Korpus ist öffentlich zugänglich.

Neben diesen Neuakquisitionen ist in DeReKo-2022-I jeweils der neue Jahrgang 2021 der fortlaufend bereitgestellten Zeitungen und Zeitschriften (insgesamt 253 Titel) in DeReKo integriert. In Summe ist DeReKo gegenüber 2021 um 2,36 Milliarden Wörter angewachsen.

5 Herausforderungen und neue Möglichkeiten der linguistischen Erschließung

5.1 Korpora ohne Forschungswerkzeuge sind wenig hilfreich

Die von DeReKo erreichte Größe von 53 Milliarden Wörtern macht jedoch auch Folgendes deutlich: Mit einem sehr großen, breit gestreuten Korpus allein ist linguistisch zunächst nichts gewonnen. Linguistische Korpora sind in der Regel zu groß und rechtlich eingeschränkt, als dass man sie einfach herunterladen könnte und – sowohl was die Daten selbst als auch ihre notwendige Kodierung betrifft – zu opak strukturiert und multidimensional, als dass man sie ohne Weiteres linguistisch interpretieren könnte. Seit etwa 2010 umschreibt der Programmbereich Korpuslinguistik am IDS seinen Ansatz zur Lösung der Herausforderung, Korpora so umfangreich wie möglich linguistisch nutzbar zu machen, frei nach Gray (2003) mit Variationen des Mottos *if the data cannot move, pave ways to put the*

9 Die zu twi21 gehörigen Twitter-IDs sind außerdem hier herunterladbar: http://corpora.ids-mannheim.de/slides/2022-03-15-DeReKo-Gesamtkonzept/twi21_ids.xz (Stand: 3.8.2022).

computation near the data (vgl. Kupietz et al. 2010; Kupietz/Diewald/Fankhauser 2018; Kupietz/Diewald/Margaretha 2022). Die Idee war damals alles andere als neu. Praktisch alle seit den 1990er Jahren entwickelten größeren synchronen Korpora waren schon aus rechtlichen Gründen nicht herunterladbar. Stattdessen gab es Suchmaschinen wie anfangs REFER (Brückner 1989), später COSMAS I (al Wadi 1994) für DeReKo und die IMS Corpus Workbench (Christ 1994), mit deren Hilfe Nutzer/-innen dann auch über eine Netzwerkverbindung in Korpora suchen konnten und z. B. KWICs als Suchresultate bekommen haben. Zumindest in der Linguistik neuer war damals die im Umfeld des Grid Computing entwickelte Verallgemeinerung des Ansatzes, nämlich aus der Not eine Tugend zu machen und anstelle der Daten grundsätzlich lieber die Computerprogramme zu ihrer Analyse zu verschicken. Ebenso wie in anderen Bereichen, wo sich dieser Ansatz nicht allgemein durchsetzen konnte, hat er auch in seiner Implementation am IDS seither einige pragmatische Änderungen erfahren.

5.2 Öffnung des Korpuszugriffs auf mehreren Ebenen

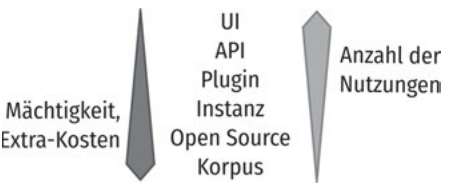

Abb. 2: Ebenen zum Zugriff DeReKo

Der aktuelle Entwurf sieht 6 Ebenen vor, auf denen Nutzer/-innen mit DeReKo bzw. KorAP interagieren können: 1) das normale User-Interface, 2) eine API zum programmatischen Zugriff, 3) Schnittstellen für eigene UI-Plugins, 4) die Möglichkeit speziell konfigurierter KorAP-Instanzen, 5) die Modifikation des Quellcodes und 6) die Möglichkeit direkt mit dem Korpus zu arbeiten. Wie Abbildung 2 veranschaulicht, haben diese grundsätzlich die Eigenschaften: je niedriger die Ebene, desto größer der Gestaltungsspielraum der Möglichkeiten, aber auch der notwendige individuelle Aufwand und entsprechend zwangsläufig desto niedriger auch die Anzahl der Nutzungen (Details siehe Kupietz/Diewald/Margaretha 2022). Die Funktionsweise des Modells beruht vor allem auf der Idee, möglichst viele Anwendungen auf einer möglichst hohen und damit für beide Seiten leicht handhabbaren Ebene zu erlauben und so das Kopieren von Daten weitgehend überflüssig zu machen und ihre Analyse methodisch und technisch übergreifend

zu unterstützen. Die Ebenen einzelner Anwendungen können dabei mit zunehmendem Funktionsumfang von KorAP und veränderlichen Anforderungen mit der Zeit wechseln.

Im Folgenden soll auf die API-Ebene näher eingegangen werden, die Nutzer/-innen bei nicht wesentlich höherem Aufwand neue Möglichkeiten mit DeReKo und etwa jetzt schon bei kontrastiven Studien zum Rumänischen und Ungarischen eröffnet.

5.3 Reproduzierbare DeReKo-Analysen mit R und Python

Um die Nutzung der erweiterten Möglichkeiten des Zugriffs auf der API-Ebene so einfach wie möglich zu gestalten, bietet KorAP Client-Bibliotheken für die Programmiersprachen R und Python an (Kupietz/Diewald/Margaretha 2020).[10] Über die API sind grundsätzlich alle KorAP-Funktionalitäten, einschließlich der Definition virtueller Korpora und komplexer Anfragen mittels aller unterstützten Anfragesprachen möglich. Die Funktionsäquivalenz ist dadurch sichergestellt, dass KorAP's Web-Benutzeroberfläche Kalamar (Diewald/Barbu Mititelu/Kupietz 2019) selbst diese APIs für die gesamte Kommunikation mit dem Backend-System verwendet. Hinsichtlich urheberrechtlicher Hürden bietet die API-Nutzung den Vorteil, dass rein quantitative Funktionen, die keine Belegstellen zurückliefern, kein registriertes Benutzerkonto erfordern und damit uneingeschränkt durch andere reproduzierbar sind. Generell ist die Reproduktion komplexer oder mehrteiliger Anfragen und entsprechender Visualisierungen ein wichtiges Anwendungsfeld für KorAP auf API-Ebene. Gleiches gilt für die Replikation von Analysen mit veränderten Korpusausschnitten, veränderten Suchausdrücken und/oder geänderten Parametern. Dies gilt auch außerhalb der Korpuslinguistik im engeren Sinne. So kann der Rat für deutsche Rechtschreibung durch den programmatischen Zugriff auf KorAP für einen neuen Beobachtungszeitpunkt einfach automatisch die Plots für alle unter Beobachtung stehenden Varianten neu erzeugen, anstatt alle Anfragen erneut manuell für den neuen Zeitraum auszuführen – was zudem auch fehleranfällig wäre. Außerdem kann er leicht seine Befunde mit denen anders definierter Korpora vergleichen.

10 Die Dokumentation zum direkten Zugriff auf die API mithilfe beliebiger Programmiersprachen ist auf dem GitHub-Wiki der KorAP-Benutzer- und Rechte-Verwaltungskomponente Kustvakt (Margaretha et al. 2015–) zu finden: https://github.com/KorAP/Kustvakt/wiki (Stand: 3.8.2022).

Ein weiteres, volatiles Anwendungsfeld für die API-Funktionalitäten sind zudem prototypische Implementierungen von Funktionen, die durch das Backend und/oder das User-Interface noch nicht vollständig unterstützt werden und so zunächst gemeinsam mit der (anwendbaren) Methodik entwickelt bzw. weiterentwickelt werden können. Dies betrifft derzeit noch die Sortierung und Aggregierung von Suchtreffern, die Kookkurrenzanalyse und die Visualisierung quantitativer Ergebnisse.

Schnelle Überprüfbarkeit von Hypothesen durch interaktive Visualisierungen

Das RKorAPClient-Paket kann in R selbst oder seiner integrierten Entwicklungsumgebung mit grafischer Benutzeroberfläche RStudio, einfach installiert werden.[11] Um seinen Einsatz möglichst niedrigschwellig zu machen, enthält das Paket über die Wrapper für die eigentlichen API-Funktionen hinaus zahlreiche Funktionen, die typische linguistische Workflows unterstützen. In Listing 2 wird z. B. das Frequenzverhältnis von dem einem Nomen (flektiert) voran- bzw. (unflektiert) nachgestellten ‚pur' über die Zeit ermittelt. Dazu wird zunächst KorAP's R-Bibliothek geladen, dann wird ein Vektor mit den beiden Anfragen und ein Vektor mit 42 virtuellen Korpora (vcs) definiert. Letztere sind alle mittels eines regulären Ausdrucks auf die Textsorten Zeitungen und Zeitschriften eingeschränkt. Außerdem ist jedes der 42 virtuellen Korpora auf ein Publikationsjahr von 1980 bis 2021 eingeschränkt. Nach der Eröffnung einer neuen Verbindung zum KorAP-Server, werden mit Hilfe der Funktion frequencyQuery die 2×42 Frequenzanfragen gestellt. Das Ergebnis, eine Tabelle mit 8 Spalten (u. a. mit den absoluten und relativen Frequenzen) und 84 Zeilen, wird dann direkt an eine ebenfalls vom RKorAPClient-Paket zur Verfügung gestellte Plot-Funktion weitergeleitet. Ein Screenshot des resultierenden interaktiven Plots (es handelt sich um eine HTML-Datei mit JavaScript) ist in Abbildung 3 dargestellt. Die Abbildung zeigt den jährlichen prozentualen Anteil der Treffer der beiden Suchausdrücke (mit Konfidenzintervallen) und bereits eine der interaktiven Funktionen: Beim Überfahren mit dem Mauszeiger werden genauere Informationen zu den jeweiligen Datenpunkten angezeigt. Außerdem lassen sich einzelne Kurven durch Klicks in die Legende ein- und ausblenden. Diese Funktionalität erweist sich als besonders hilfreich bei einer größeren Anzahl an Suchausdrücken, da sich Nutzer/-innen dadurch eine Auswahl von Kurven interaktiv in der Grafik zusammenstel-

11 In R: install.packages(„RKorAPClient") – in RStudio: Tools → Install Packages → RKorAPClient.

len können, die sie fokussiert gemeinsam (oder auch einzeln ausgewählt) mit angepasster Skalierung betrachten können. Methodisch am relevantesten ist aber die Funktionalität, dass das Anklicken eines Datenpunktes ein neues Browserfenster mit genau der dem Datenpunkt zugrundeliegenden KorAP-Anfrage öffnet. So können die aggregierten quantitativen Ergebnisse schnell z. B. auf falsch Positive überprüft werden und generell quantitative Analysen und qualitative Interpretationen eng miteinander verknüpft werden (vgl. Kupietz et al. 2017, S. 326 f., Perkuhn/Kupietz 2018, S. 86 f.).

Listing 1: Vollständiger R-Code zur Erzeugung von Abbildung 3

```
library(RKorAPClient)

anfragen <- c("[tt/l=pur] [tt/p=NN]",
              "[tt/p=NN] pur")

vcs <- paste("textType = /Zeit.*/ & pubDate in ", c(1980:2021))

new("KorAPConnection", verbose=T) %>%
  frequencyQuery(anfragen, vcs, as.alternatives = TRUE) %>%
  hc_freq_by_year_ci(as.alternatives = TRUE)
```

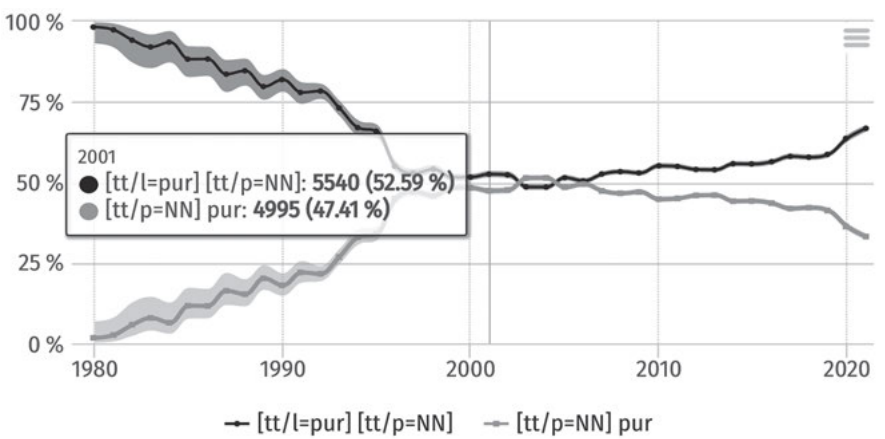

Abb. 3: Frequenzverhältnis zwischen voran- und nachgestelltem ‚pur' in DeReKo-Presseerzeugnissen zwischen 1980 und 2021; Screenshot des mit Listing 1 erzeugten interaktiven Plots

5.4 Kontrastive Linguistik

Die 2013 vom IDS und den Akademien in Polen, Rumänien und Ungarn gegründete offene Initiative European Reference Corpus EuReCo (Kupietz et al. 2017, 2020; Trawiński/Kupietz 2021) verfolgt im Wesentlichen zwei Ziele: 1) die Kräfte für die linguistische Erschließung sehr großer Korpora durch Forschungssoftware zu bündeln und 2) die Möglichkeiten kontrastiv linguistischer Forschung auf Grundlage vergleichbarer Korpora zu verbessern. Grundidee ist dabei, die Voraussetzungen für dynamisch definierbare vergleichbare Korpora auf Basis vorhandener großer Korpora zu schaffen, mit 1) hoher linguistischer Qualität und breiter Einsetzbarkeit, 2) dynamisch anpassbaren und optimierbaren Vergleichskriterien, bei 3) realistischem Aufwand. Im EuReCo-Kontext sind zurzeit neben DeReKo das Referenzkorpus der Rumänischen Gegenwartssprache CoRoLa (Tufiș et al. 2016) und seit 2021 das vollständige Ungarische Nationalkorpus HNC (Váradi 2002) über KorAP abfragbar.[12] Von der Möglichkeit zur Konstruktion eigener oder der Verwendung vordefinierter[13] virtueller vergleichbarer Korpora abgesehen, bringt bereits die Verfügbarkeit über eine einheitliche Analyseplattform eine Vereinfachung kontrastiver Studien mit sich.

Methodisch neue Möglichkeiten eröffnet die in KorAP's Client-Bibliotheken integrierte Kookkurrenzanalyse-Funktion (FVG) – etwa zur sprachvergleichenden Untersuchung von Funktionsverbgefügen, die in ersten deutsch-rumänischen Studien auch im Hinblick auf den Einfluss von Korpusvergleichbarkeit und Korpuszusammensetzung durchgeführt wurde (Kupietz/Trawiński im Ersch.). Bei den wenigen bisher untersuchten FVG hat sich u. a. gezeigt, dass die Verlinkung von Ergebnissen der Kookkurrenzanalyse mit Suchanfragen, die die zugrundeliegenden Konkordanzen anzeigen, gerade im Sprachvergleich hilfreich sind, um Artefakte (z. B. auch aufgrund partieller Text-Dubletten) bzw. falsch positive Treffer zu identifizieren. Die dynamische Anpassbarkeit der virtuellen Korpora hat sich außerdem als hilfreich erwiesen, um aus unterschiedlichen Zusammensetzungen bzgl. Textsorten und thematischer Domäne resultierende Effekte, z. B. durch einfaches Ausprobieren anderer Zusammensetzungen, als Artefakte zu identifizieren und zu dämpfen. Die vorläufigen Ergebnisse deuten darauf hin, dass mehr noch, als das bei einzelsprachlichen Untersuchungen der Fall ist, die Kompositionsprinzipien virtueller vergleichbarer Korpora stark mit

12 Siehe https://korap.racai.ro/ (Stand: 3.8.2022) bzw. https://korap.nlp.nytud.hu (Stand: 3.8.2022).

13 Bisher nur Deutsch-Rumänisch.

der Fragestellung variieren werden, da jeweils andere und im Detail schwer vorhersagbare Vergleichbarkeitskriterien relevant sind.

Ähnliche experimentelle Studien werden derzeit zum Deutsch-Ungarischen Vergleich in engem Zusammenhang mit der Weiterentwicklung von KorAP's Kookkurrenzanalysefunktionalitäten durchgeführt. Außerdem ist zur EuReCo-Erweiterung die Überführung des Polnischen Nationalkorpus (Przepiórkowski et al. 2010) in das KorAP-XML-Format in Arbeit.

6 Resümee

Die Weiterentwicklung von DeReKo ist eingebettet in ein Gesamtkonzept mit den Zielen erstens weiterhin eine stetige und verlässliche Forschungsdatengrundlage anzubieten und zweitens das Potenzial dieser weiterhin optimal linguistisch erschließbar zu machen – sowohl für eine breite Nutzung als auch für spezielle und anspruchsvolle Anwendungen, getreu dem Prinzip, dass Einfaches einfach und Komplexes möglich sein sollte. Die Mittel zum Erreichen dieser Ziele sind ein in dieser Hinsicht optimierter Einsatz der vorhandenen Ressourcen und die Öffnung aller ohnehin vorhandenen Schnittstellen (im Rahmen der rechtlich eingeräumten Nutzungsmöglichkeiten).

Bezüglich der Forschungsdatengewinnung und -aufbereitung haben wir insbesondere versucht zu zeigen, dass große und breit gestreute Korpora wie DeReKo die Anwendung heuristischer Verfahren unabdingbar machen. Für die Verwendungsseite hat dies zur Konsequenz, dass solche Korpora grundsätzlich nicht mehr isoliert von ihrer Entstehung betrachtet werden können. Für die Seite der Aufbereitung hat das zur Folge, dass Genauigkeit als Evaluationskriterium durch eine Reihe weiterer Optimierungskriterien ergänzt werden muss, deren Abwägung eine enge Verknüpfung linguistischer, informatischer, softwaretechnischer und infrastruktureller Kompetenzen erfordert.

Trotz dieser Herausforderungen und dank einiger vorgestellter Strategien zum Umgang mit Fehlern und Unzulänglichkeiten konnte DeReKo 2021 um 2,4 Milliarden Wörter erweitert und bzgl. seiner Abdeckung in den Bereichen Internetbasierte Kommunikation und Fachsprache(n) verbessert werden. Neue Möglichkeiten zur Nutzung von DeReKo eröffnen außerdem die KorAP-Client-Bibliotheken für R und Python. Sie erleichtern die Reproduzierbarkeit und Replizierbarkeit von Korpusanalysen und ermöglichen die schnelle Abduzierbarkeit und Überprüfbarkeit von Hypothesen und eine enge Verbindung quantitativer Analysen mit qualitativen Interpretationen durch interaktive Visualisierungen.

Die Erweiterung von EuReCo um das vollständige Ungarische Nationalkorpus Ende 2021 erweitert außerdem das Spektrum sprachvergleichender Forschungspotenziale im Kontext von DeReKo und trägt durch die Vergrößerung der Nutzer- und Entwickler/-innenbasis zusätzlich auch indirekt methodisch, ökonomisch und infrastrukturell zur Verbesserung der linguistischen Nutzungsmöglichkeiten von DeReKo und anderen sehr großen Korpora bei.

Literatur

al Wadi, Doris (1994): COSMAS – Ein Computersystem für den Zugriff auf Textkorpora. Mannheim: Institut für Deutsche Sprache.

Bański, Piotr/Fischer, Peter M./Frick, Elena/Ketzan, Erik/Kupietz, Marc/Schnober, Carsten/Schonefeld, Oliver/Witt, Andreas (2012): The new IDS corpus analysis platform: challenges and prospects. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Doğan, Mehmet Uğur/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the eighth international conference on language resources and evaluation (LREC 2012). Paris: European Language Resources Association (ELRA), S. 2905–2911.

Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lüngen, Harald (2011): The morphosyntactic annotation of DeReKo: interpretation, opportunities, and pitfalls. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich Hermann (Hg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.–24.9.2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen: Narr, S. 451–469.

BNC Consortium (2007): The British National Corpus, XML Edition, Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554 (Stand: 3.8.2022).

Bodmer, Franck (1996): Aspekte Der Abfragekomponononente von COSMAS-II. In: LDV-INFO. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung 8, S. 112–122.

Brückner, Tobias (1989): REFER. Benutzerhandbuch. Mannheim: Institut für Deutsche Sprache.

Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the seventh conference on international language resources and evaluation (LREC'10). Paris: European Language Resources Association (ELRA).

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX´94. 3rd conference on Computational Lexicography and text research. Budapest, S. 22–32.

Cotgrove, Louis A. (2022): #GlockeAktiv: A Corpus Linguistic investigation of German online youth language. PhD Thesis. Nottingham: University of Nottingham.

Cotgrove, Louis A. (im Dr.): New opportunities for researching digital youth language: the NottDeuYTSch corpus. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr.

Diewald, Nils (2016–): https://github.com/KorAP/KorAP-XML-Krill (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.6452005 (Stand: 3.8.2022).

Diewald, Nils (2022): Matrix and double-array representations for efficient finite state tokenization. In: Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Kupietz, Marc/Lüngen, Harald (Hg.): Proceedings of the LREC 2022. Workshop on Challenges in the Management of Large Corpora (CMLC-10 2022) Marseille: European Language Resources Association (ELRA), 2022, S. 20–26.

Diewald, Nils/Barbu Mititelu, Verginica/Kupietz, Marc (2019): The KorAP user interface. Accessing CoRoLa via KorAP. In: Revue Roumaine de Linguistique. On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo64, 3, S. 265–277.

Diewald, Nils/Kupietz, Marc/Lüngen, Harald (2022): Tokenizing on scale: Preprocessing large text corpora on the lexical and sentence level. In: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (Hg.): Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim. Mannheim: IDS-Verlag.

Diewald, Nils/Margaretha, Eliza/Kupietz, Marc (2021): Lessons learned in quality management for online research software tools in Linguistics. In: Lüngen, Harald/Kupietz, Marc/Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Pisetta, Ines (Hg.): Proceedings of the workshop on challenges in the management of large corpora (CMLC-9). (Online-Event). Mannheim: Leibniz-Institut für Deutsche Sprache, S. 20–26.

Gîfu, Daniela/Moruz, Alex/Bolea, Cecilia/Bibiri, Anca/Mitrofan, Maria (2019): The methodology of building CoRoLa. In: Revue Roumaine de Linquistique. On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo 64, 3, S. 241–253.

Gray, Jim (2003): Distributed Computing Economics. Technical Report MSR-TR-2003-24, Microsoft Research. Redmond, WA: Microsoft Corporation, One Microsoft Way.

Hall, Johan/Nivre, Joakim (2008): A dependency-driven parser for German dependency and constituency representations. In: Proceedings of the workshop on parsing German. Columbus, Ohio. Association for Computational Linguistics, S. 47–54. https://aclanthology.org/W08-1007 (Stand: 3.8.2022).

Harders, Peter/Diewald, Nils/Kupietz, Marc/Schnober, Carsten (2020–): https://github.com/KorAP/KorAP-XML-TEI (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.6451963 (Stand: 3.8.2022).

Kamocki, Paweł/Hannesschläger, Vanessa/Hoorn, Esther/Kelli, Aleksei/Kupietz, Marc/Lindén, Krister/Puksas, Andrius (2021): Legal issues related to the use of twitter data in language research. In: Monachini, Monica/Eskevich, Maria (Hg.): Proceedings of CLARIN annual conference. 27 – 29 September 2021, virtual edition. Utrecht: CLARIN, S. 150–153.

Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28, S. 71–97.

Kupietz, Marc (2005): Near-duplicate detection in the IDS corpora of written German (Technical report IDS-KT-2006-01). Mannheim: Institut für Deutsche Sprache.

Kupietz, Marc (2015): Constructing a corpus. In: Durkin, Philip (Hg.): The Oxford handbook of Lexicography. (= Oxford Handbooks in Linguistics). Oxford: Oxford University Press, S. 62–75.

Kupietz, Marc/Diewald, Nils (2022): KorAP-Tokenizer. https://github.com/KorAP/KorAP-Tokenizer (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.5862064 (Stand: 3.8.2022).

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji. (Hg.): Working papers in corpus-based Linguistics and language education. Bd. 3. Tokyo: University of Foreign Studies (TUFS), S. 53–59.

Kupietz, Marc/Schmidt, Thomas (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In: Eichinger, Ludwig M. (Hg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. (= Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/München/Boston: De Gruyter, S. 297–322. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-34824 (Stand: 9.8.2022).

Kupietz, Marc/Trawiński, Beata (im Ersch.): Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo. In: Akten des XIV. Kongresses der Internationalen Vereinigung für Germanische Sprach- und Literaturwissenschaft (IVG). Berlin u. a.: Lang.

Kupietz, Marc/Diewald, Nils/Fankhauser, Peter (2018): How to get the computation near the data: improving data accessibility to, and reusability of analysis functions in corpus query platforms. In: Bański, Piotr/Kupietz, Marc/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Witt, Andreas (Hg.): Proceedings of the LREC 2018 workshop "Challenges in the management of large corpora (CMLC-6)", 07 May 2018 – Miyazaki, Japan. Paris: European language resources association (ELRA), S. 20–25.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2020): RKorAPClient: an R package for accessing the German Reference Corpus DeReKo via KorAP. In: Calzolari, Nicoletta/Béchet, Frédéric/Blache, Philippe/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 12th international conference on language resources and evaluation (LREC), May 11–16, 2020, Palais du Pharo, Marseille, France. Paris: European Language Resources Association (ELRA), S. 7016–7021.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2022): Building paths to corpus data: A multi-level least effort and maximum return approach. In: Fišer, Darja/Witt, Andreas (Hg.): CLARIN. The infrastructure for language resources. (= Digital Linguistics 1). Berlin: De Gruyter.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: a primordial sample for linguistic research. In: Calzolari/Choukri/Maegaard/Mariani/Odijk/Piperidis/Rosner/Tapias (Hg.), S. 1848–1854.

Kupietz, Marc/Diewald, Nils/Hanl, Michael/Margaretha, Eliza (2017): Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In: Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung. Proceedings of the Methodenmesse im Rahmen der Jahrestagung des Instituts für Deutsche Sprache. 9. März 2016, Mannheim. (= Jahrbuch des Instituts für Deutsche Sprache 2016). Berlin/Boston: De Gruyter, S. 319–329

Kupietz, Marc/Witt, Andreas/Bański, Piotr/Tufiş, Dan/Cristea, Dan/Váradi, Tamás (2017): EuReCo - Joining forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (Hg.): Proceedings of the workshop on challenges in the management of large corpora and big data and natural language processing (CMLC-5+BigNLP) 2017

including the papers from the web-as-corpus (WAC-XI) guest section. Birmingham, 24 July 2017. Mannheim: Leibniz-Institut für Deutsche Sprache, S. 15–19.

Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, Sylviane/Lefer, Marie-Aude (Hg.): Translating and comparing languages: corpus-based insights. Selected proceedings of the fifth using corpora in contrastive and trranslation studies conference. (= Corpora and Language in Use. Proceedings 6). Louvain-la-Neuve: Presses universitaires de Louvain, S. 257–273.

Laney, Douglas (2001): 3D data management: controlling data volume, velocity, and variety. (= Application Delivery Strategies 949). META Group. https://www.bibsonomy.org/bibtex/742811cb00b303261f79a98e9b80bf49 (Stand: 9.8.2022).

Leibniz-Institut für Deutsche Sprache (2020): Satzung des Leibniz-Instituts für Deutsche Sprache (IDS). Fassung vom 18.5.2020. https://www.ids-mannheim.de/fileadmin/org/pdf/IDS_Satzung_2020-05-18.pdf (Stand: 9.8.2022).

Liu, Jianzheng/Li, Jie/Li, Weifeng/Wu, Jiansheng (2016): Rethinking big data: A review on the data quality and usage issues. In: ISPRS Journal of Photogrammetry and Remote Sensing 115, S. 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006 (Stand: 9.8.2022).

Lüngen, Harald/Kupietz Marc (2020): IBK- und Social Media-Korpora am Leibniz-Institut für Deutsche Sprache. In: Marx, Konstanze/Lobin, Henning/Schmidt, Axel (Hg.): Deutsch in Sozialen Medien. Interaktiv – multimodal – vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin/Boston: De Gruyter, S. 319–344.

Lüngen, Harald/Sperberg-McQueen, C. Michael (2012): A TEI P5 document grammar for the IDS text model. In: Journal of the Text Encoding Initiative 3. https://journals.openedition.org/jtei/pdf/508 (Stand: 9.8.2022).

Manning, Christopher D./Surdeanu, Mihai/Bauer, John/Finkel, Jenny/Bethard, Steven J./McClosky, David (2014): The Stanford CoreNLP natural language processing toolkit. In: Bontcheva, Kalina/Zhu, Jingbo (Hg.): Proceedings of 52nd annual meeting of the association for Computational Linguistics: system demonstrations. Association for Computational Linguistics, S. 55–60.

Margaretha, Eliza/Hanl, Michael/Diewald, Nils/Kupietz, Marc/Bodmer, Franck (2015–): https://github.com/KorAP/Kustvakt (Stand: 9.8.2022). https://doi.org/10.5281/zenodo.5026507 (Stand: 9.8.2022).

Müller, Thomas/Schmid, Helmut/Schütze, Hinrich (2013): Efficient higher-order CRFs for morphological tagging. In: Yarowsky, David/Baldwin, Timothy/Korhonen, Anna/Livescu, Karen/Bethard, Steven (Hg.): Proceedings of the 2013 conference on empirical methods in natural language processing. Seattle, Washington, USA, October 2013, S. 322–332. https://aclanthology.org/D13-1032 (Stand: 9.8.2022).

Perkuhn, Rainer/Kupietz, Marc (2018): Visualisierung als aufmerksamkeitsleitendes Instrument bei der Analyse sehr großer Korpora. In: Bubenhofer, Noah/Kupietz, Marc (Hg.): Visualisierung sprachlicher Daten. Visual Linguistics – Praxis – Tools. Heidelberg: Heidelberg University Publishing, S. 63–90.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= UTB 3433). Paderborn: Fink.

Przepiórkowski, Adam/Górski, Rafał L./Łaziński, Marek/Pęzik, Piotr (2010): Recent developments in the National Corpus of Polish. In: Calzolari/Choukri/Declerck/Maegaard/Mariani/Odijk/Piperidis/Rosner/Tapias (Hg.), S. 994–997.

Schirrmeister, Lars/Rummel, Marlene/Heine, Antje/Suppus, Nina/Mendoza Sánchez, Bárbara (2021): Gingko – ein Korpus der ingenieurwissenschaftlichen Sprache. In: Deutsch als Fremdsprache 4, 214–224. doi.org/10.37307/j.2198-2430.2021.04.04 (Stand: 9.8.2022).

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing. Manchester, United Kingdom, September 1994.

Trawiński, Beata/Kupietz, Marc (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, Henning/Wöllstein, Angelika/Witt, Andreas (Hg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston: De Gruyter, S. 209–234.

Tufiş, Dan/Barbu Mititelu, Verginica/Irimia, Elena/Dumitrescu, Ştefan D./Boroş, Tiberiu (2016): The IPR-cleared corpus of Contemporary written and spoken Romanian language. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Goggi, Sara/Grobelnik, Marko / Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the tenth international conference on language resources and evaluation (LREC'16). Paris/Portoroz: European Language Resources Association (ELRA) S. 2516–2521.

Váradi, Tamás (2002): The Hungarian National Corpus. In: González Rodríguez, Manuel/Suárez Araujo, Carmen (Hg.): Proceedings of the third international conference on language resources and evaluation (LREC 2002). Las Palmas/Paris: European Language Resources Association (ELRA), S. 385–389.