Marc Kupietz (Mannheim)/Nils Diewald (Mannheim)/
Michael Hanl (Mannheim)/Eliza Margaretha (Mannheim)

Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP

Abstract: KorAP, die neue Korpusanalyseplattform des IDS, die COSMAS II im Laufe der kommenden 2–3 Jahre ablösen wird, bietet gerade zur Erforschung grammatischer Variation einige besondere Funktionalitäten. Grundlegend ist beispielsweise, dass KorAP die Repräsentation und Abfrage beliebiger und beliebig vieler Annotationsschichten, zum Beispiel zu Konstituenz- und Dependenzrelationen, unterstützt und damit die Suche nach speziellen grammatischen Phänomenen erleichtert oder erst möglich macht. Darüber hinaus unterstützt KorAP die Konstruktion virtueller Korpora anhand von Metadatenvariablen und erleichtert damit kontrastive Untersuchungen. Der vorliegende Artikel erläutert die für die grammatische Variationsforschung relevanten KorAP-Funktionalitäten im Einzelnen und gibt einen Einblick in ihre Grundlagen.

1 Einleitung

Im Laufe der nächsten zwei bis drei Jahre wird KorAP COSMAS II als primäre Analyseplattform für das Deutsche Referenzkorpus (DeReKo; Kupietz et al. 2010; Kupietz/Lüngen 2014) sukzessive ablösen. Die neue Plattform wird dabei zahlreiche Funktionen insbesondere zur Untersuchung grammatischer Phänomene und deren Variation bieten. Hierzu gehört beispielsweise die Möglichkeit zur Repräsentation und Abfrage von Konstituenz- und Dependenzannotationen, aber auch die prinzipielle Unterstützung paralleler Aufrufe einer unbegrenzten Anzahl verschiedener, potenziell konkurrierender Annotationsschichten.

Relevant für die grammatische Variationsforschung ist außerdem KorAPs Unterstützung prinzipiell unbeschränkt großer Korpora. Nicht nur weil DeReKo eine Größe von mittlerweile 30 Milliarden Wörtern mit einer Wachstumsrate von über einer Milliarde Wörter pro Jahr erreicht hat, sondern auch weil Korpuspartitionen, die durch Kombination metasprachlicher Variablenrestriktionen erzeugt wurden (z.B. *1995 in Bayern veröffentlichte Zeitungskommentare*), schnell so klein werden, dass selbst bei der Untersuchung häufiger Phänomene in Abhängigkeit von verschiedenen Variationsalternativen wie Zeitraum, Textsorte oder

Region kaum noch ausreichend Belege gefunden werden können, um diese zu interpretieren.

Im Folgenden wird eine Auswahl weiterer für die Variationsforschung relevanter Funktionalitäten genauer erläutert.

2 Multiple Anfragesprachen

KorAP wurde entwickelt, um COSMAS II als primären Zugang zu DeREKo abzulösen. Entsprechend war die Unterstützung der COSMAS-II-Anfragesprache (Bodmer 1996) ein Hauptkriterium, um Nutzern des Vorgängersystems den Umstieg so leicht wie möglich zu gestalten. Gleichzeitig sollten erweiterte Analysekonstrukte unterstützt werden, die von anderen etablierten Anfragesprachen bereits erfasst werden (z.B. die Analyse von Abhängigkeitsstrukturen in Annis-QL; Rosenfeld 2010).

Um die Entwicklung einer neuen, mächtigeren Anfragesprache zu vermeiden, die insbesondere bei sehr einfachen Anfragen oft als Hürde wahrgenommen wird, wurde für KorAP eine Unterstützung multipler Anfragesprachen beschlossen.

Aus Nutzersicht hat dies zahlreiche Vorteile:

- Nutzer anderer Korpusanalyseplattformen (Annis, SketchEngine etc.) müssen keine weitere Sprache erlernen;
- Nutzer können ihre Forschungsergebnisse ohne Reformulierung über Korpusanalyseplattformen hinweg vergleichen (vgl. Abb. 1);
- Nutzer, die sich auf die Erforschung einzelner Sprachphänomene konzentrieren, können die für sie adäquateste Anfragesprache wählen.

Aus technischer Sicht stellt die Unterstützung multipler Anfragesprachen jedoch eine Herausforderung dar, da unterschiedliche Anfragekonstrukte unterschiedliche Anforderungen an das Analysesystem stellen. Aus diesem Grund wurden im Rahmen des KorAP-Projekts zwei Initiativen gestartet: zum einen die Erarbeitung einer Spezifikation zur Beschreibung und zum Vergleich von Eigenschaften verschiedener Korpus-Anfragesprachen (CQLF; Bański/Frick/Witt 2016), zum anderen die Entwicklung eines erweiterbaren Anfrageprotokolls zur Vereinheitlichung multipler Anfragesprachen (*KoralQuery*; Bingel/Diewald 2015).

Das in KorAP eingesetzte *KoralQuery* ist ein auf JSON-LD (Sporny et al. 2014) basierendes Protokollformat zur Serialisierung von Korpus-Anfragen. Dabei werden Anfragen verschiedener Sprachen mittels eines Konverters (*Koral*; Bingel 2015) in einen abstrakten Syntax-Baum transformiert und dieser wiederum in ein *KoralQuery*-Dokument übersetzt (siehe Abb. 2).

Abb. 1: Identische Suchanfrage formuliert in COSMAS II QL und in Poliqarp

Abb. 2: 2-Schritt-Transformation beliebiger Anfragen beliebiger Korpus-Anfragesprachen in KoralQuery

Weitere Prozesse in KorAP verarbeiten nach der Konvertierung ausschließlich das vereinheitlichte KoralQuery-Format. Hauptaugenmerk bei der Entwicklung des Protokolls wurde auf die Neutralität in Bezug auf Anfragesprachen, Forschungsfragen und Korpusdaten gelegt, außerdem auf die Erweiterbarkeit und Abwärtskompatibilität, um zukünftig weitere Anfragesprachen unterstützen zu können, deren Anfragekonstrukte bisher noch keine Berücksichtigung fanden. Koral unterstützt derzeit COSMAS II QL, Annis, Poliqarp (eine Variante des bekannten CQP), CLARIN Federated Content Search Anfragesprachen und eine Teilmenge von CQL.

Um die in Abschnitt 3 erläuterten multiplen Annotationsschichten zu unterstützen, wurden die Anfragesprachen um die Möglichkeit erweitert, für erfragte Annotationen die jeweilige Ressource zu definieren. Dabei erhalten alle Anno-

tationsanfragen die Struktur `Foundry/Layer=Annotation`. Hierbei steht *Foundry* für die Kennung einer beliebigen Ressource von Annotationsdaten (z.B. `tt` für Annotationen, die mit dem TreeTagger-Tool durchgeführt wurden, oder `dereko` für Strukturinformationen aus den DEREKO-Quellen), während *Layer* für die Kennung einer beliebigen Annotationsebene steht (z.B. `p` für Part-of-Speech-Annotationen, oder `s` für Strukturannotationen). Beispielsweise kennzeichnet `tt/p=NN` ein von TreeTagger identifiziertes normales Nomen der Part-of-Speech-Annotationsebene. Dieser Mechanismus wurde sprachenübergreifend implementiert. Für Anfragen ohne Ressourcen-Definition werden Standardquellen angenommen, die vom System vorgegeben und vom Nutzer überschrieben werden können.

3 Multiple Annotationsschichten

Die Unterstützung einer prinzipiell unbegrenzten Anzahl beliebiger linguistischer Annotationen war einer der Hauptgründe für die Entwicklung von KorAP. Das Ziel bezüglich DEREKO war und ist dabei, alle Annotationen, an denen von Nutzer- oder Tool-Anbieterseite aus Interesse besteht, DEREKO hinzuzufügen und über KorAP abfragbar zu machen. Zum einen sollte so gegenüber COSMAS II das Spektrum abfragbarer Annotationsebenen etwa um Abhängigkeits-, Konstituents-, Koreferenz-, RST-Relationen etc. erweitert werden. Zum anderen sollte auch die Menge unterschiedlicher, konkurrierender Annotationen innerhalb einer Ebene leichter erweiterbar sein, um den Umgang mit unsicheren Klassifikationen zu verbessern.

Intern ist KorAP in der Lage, Annotationen auf Basis von Tokens (z.B. Lemma, Part-of-Speech, Named-Entity-Klassen), Token-Sequenzen (z.B. Konstituenten, Sätzen, Dokumentstrukturen) und als Relationen (z.B. Abhängigkeitsrelationen, anaphorische Relationen) zu repräsentieren – jeweils mit der Möglichkeit, diesen elementaren Annotationen weitere Attribute hinzuzufügen.

Durch die Kombination von Suchoperatoren ist es möglich, Abhängigkeiten verschiedener Annotationen und Oberflächenformen untereinander zu definieren und so komplexe Suchanfragen zu formulieren.

```
A) [mate/m=number:pl & mate/p=ART] {1:[cnx/p=A]? [tt/
l=Baum] }
```

Beispiel A und Abbildung 3 zeigen eine solche Suche mit Poliqarp nach einer Sequenz von Wörtern, die mit einem von MATE als Artikel im Plural identifizier-

ten Wort beginnt, welches gefolgt wird von einem optionalen, von Connexor MPT annotierten Adjektiv und einem Wort, das von TreeTagger als „Baum“ lemmatisiert wurde. Zur besseren Lesbarkeit der Ergebnisse können Teile der Suchanfrage hervorgehoben werden (wie hier die Sequenz nach dem ersten Wort).

Die Verwendung linguistischer Annotationen in Suchanfragen ist für die Untersuchung grammatischer Variationen in der Regel unverzichtbar. Wenn diese über das Finden von Beispielen hinaus für quantitative Untersuchungen, zur korpusbasierten Überprüfung von Hypothesen oder für die Exploration von Phänomenen gebraucht werden, ist ihre Verwendung jedoch nicht trivial, und es sind einige Fallstricke zu beachten, da sie typischerweise nicht den Status von Beobachtungen, sondern den theorie- und – im Fall von DEREKO – implementationsabhängiger und fehlerbehafteter Interpretationen haben (Belica et al. 2011). Dieser Umstand ist schon für Wortartannotationen relevant, für die eine Genauigkeit von bis zu 97–98% angegeben wird (siehe Giesbrecht/Evert 2009). Das Problem selbst bei solch geringen Fehlerquoten ist, dass Fehler nicht gleichmäßig und tendenziell ungünstig verteilt sind (vgl. Meurers/Müller 2009, S. 923). Zum einen sind für viele linguistische Fragestellungen geringere Genauigkeitswerte zu erwarten, da sich die Untersuchungen typischerweise auf weniger eindeutige oder wahrscheinliche Phänomene beziehen. Zum anderen kann auf Grund der in dieser Hinsicht nicht normalverteilten Fehler aus einer durchschnittlichen Genauigkeit nicht direkt auf einen Erwartungswert für die Genauigkeit der Ergebnisse zu einer bestimmten Suchanfrage geschlossen werden. Vielmehr muss einkalkuliert werden, dass zwar die meisten Konstruktionen fast immer richtig annotiert sind, andere aber fast immer falsch, so dass bei Suchanfragen zu letzteren mit einem sehr geringen Recall gerechnet werden muss.¹ Da falsch-negative Treffer nicht sichtbar und damit nicht ohne Weiteres erkennbar sind, kann eine solche Situation leicht zu Fehlschlüssen verleiten. Bei der Untersuchung grammatischer Phänomene kommt hinzu, dass bei der Suche nach Wortkombinationen anhand von Annotationen sich die Unsicherheiten multiplizieren und dass auf höheren Annotationsebenen, wie etwa bei Abhängigkeits- und insbesondere Konstituenzrelationen, entsprechend mit deutlich geringeren Genauigkeiten zu rechnen ist.²

1 Die Suche nach dem Lemma *werden* gefolgt von *sein*, annotiert als Pronomen, liefert auf den Connexor-MPT-Annotationen im virtuellen Korpus TAGGED-C von eigentlich 8945 nur 13 korrekte Treffer, was einem Recall von 0,13% entspricht.

2 Die Stanford-NLP-Gruppe gibt für den zurzeit für DEREKO verwendeten Shift-Reduce-Konstituenzparser aus den CoreNLP-Tools etwa einen F1-Score von 0,74 an. Siehe <http://nlp.stanford.edu/software/srparser.shtml>.

KorAP [mate/m=number:pl & mate/p=ART] {1:[cnx/p=A]? [tt/l=Baum]}

author contains goethe
or author contains schiller

in one collection - with Poliqarp

14 matches

Einfassung macht, **die hohen Bäume**, die den Platz rings umher bedecken, die Kühle des Orts; das hat alles so was Anzügliches, was Schauerliches. es vergeht kein Tag, immer heilig unter **den Bäumen**. Ich sage dir, dem Schulmeister standen die Tränen in den Augen, da wir gestern davon redeten, daß sie abgehauen worden - abgehauen! unrein und dumpling, **die Bäume** nehmen ihr das Tageslicht, und wenn die Nüsse reif sind, so werfen die Knaben mit Steinen darnach, und das fällt ihr auf die Nerven, das stört die Pfarrhofs, wo **die Bäume** standen, und verkaufte sie an den Meistbietenden. sie liegen! o, wenn ich Fürst wäre! Ich wollte die Pfarrerin, den Schulzen und die Kammerd Mord verwandelt. **die starken Bäume** standen ohne Laub und bereift, die schönen Hecken, die sich über die niedrige Kirchhofmauer wölbten, waren entblättert, und die Golliten am Fuß unter **den Bäumen** halten, und man wollte bis zu einem gewissen Punkte gelangen, wo ein vorstehender mächtiger Fels einen Flächenraum darbot, von wo man jetzt fangen erst **die Bäume**, die Felsen, ja Rom selbst an, mir lieb zu werden; bisher hab' ich sie immer nur als fremd gefühlt; dagegen freuten mich geringe Gegenstände als im Winter, weil **die grünen Bäume** und Hecken diesen großen Ebenen eine anmutige Verschiedenheit geben. wir fanden uns kurz vor der Abenddämmerung in Mitte der ... Ich schreibe dir später, um nur etwas zu schreiben. diese Woche hab' ich sehr vergnügt zugebracht. es wollte die vorige Woche nicht gehen, weder mit einer noch andrer Arbeit, und da es am Montage so schön Wetter war und meine Kenntnis des Himmels mich gute Tage hoffen ließ, machte ich mich mit Kaysern und meinem zweiten Fritz auf die Beine und durchging von Dienstag bis heute abend die Plätze, die ich schon kannte, und verschiedene Seiten, die ich noch nicht kannte. Dienstag abend erreichten wir Frascati, Mittwoch besuchten wir die schönsten Villen und besonders den köstlichen Antinous auf Monte Dragone. Donnerstag gingen wir von Frascati auf Monte Cavo über Rocca di Papa, wovon du einmal Zeichnungen haben sollst, denn Worte und Beschreibungen sind nichts; dann nach Albano herunter. Freitag schied Kayser von uns, dem es nicht ganz wohl war, und ich ging mit Fritz dem zweiten auf Anicia, Genzano, am See von Nemi her wieder auf Albano zurück. heute sind wir auf Castel Gandolfo und Marino gegangen und von da nach Rom zurück. das Wetter hat uns unglaublich begünstigt, es war fast das schönste Wetter des ganzen Jahrs. außer **den immergrünen Bäumen** haben noch einige Eichen ihr Laub, auch junge Kastanien noch das Laub, wenigleich gelb, es sind Töne in der Landschaft von der größten Schönheit, und die herrlichen großen Formen im nächtlichen Dunkel! Ich habe große Freude gehabt, die ich dir in der Ferne mitteile. Ich war sehr vergnügt und wohl...

Foundry	Layer	außer	den	immergrünen	Bäumen	haben	noch	einige	Eichen	ihr	Laub	auc
cnx	l	ausser	die	immergrünen	baum	haben	noch	einige	eiche	ihr	laub	au
cnx	m				PL	IND PRES			PL			
cnx	p	PREP	DET	A	N	V	ADV	PRON	N	PRON	N	AC
cnx	syn	PREMARK	PREMOD	PREMOD	PREMOD	MAIN	ADVL	PREMOD	NH	PREMOD	NH	AC
corenlp	p		ART			VAFIN	ADV	PTS	NE	PPOSAT	NN	AC
glemm	l	__ausser	__d-	__immergrün	__baum __bäumen	__haben	__noch	__einig __einig __einigen	__+E __Eiche __Eichen __eichen	__ihr __sie/ihr	__Laub	__a
mate	l	außer	der	immergrün	baum	haben	noch	einiger	eiche	ihr	laub	au
mate	m		case:dat gender:masc number:pl	case:dat degree:pos gender:masc number:pl	case:dat gender:masc number:pl	mood:ind number:pl person:3 tense:pres		case:acc gender:fem number:pl	case:acc gender:fem number:pl	case:acc gender:neut number:sg	case:acc gender:neut number:sg	
mate	p	APPR	ART	ADJA	NN	VAFIN	ADV	PIAT	NN	PPOSAT	NN	AC
opennlp	p	APPR	ART	ADJA	NN	VAFIN	ADV	PIAT	NN	PPOSAT	NN	AC
tt	l	außer	die	immergrün	Baum	haben	noch	einige	Eiche	ihr	Laub	au
tt	p	APPR	ART	ADJA	NN	VAFIN	ADV	PIAT	NN	PPOSAT	NE	AC
xlp	l	außer	der	immergrün	Baum	haben	noch	einige	eichen	ihr	Laub	au
xlp	p	PREP	DET	ADJ	NOUN	VERB	ADV	DET	NOUN	DET	NOUN	AC
xlp	c											

Add tree view

Zweiter römischer Aufenthalt by Goethe, Johann Wolfgang von (1982)

abgelangt, ihr unter **den hohen Bäumen** am ländlichen Tische gegenübersaß, die freundliche Müllerin nach Milch, der bewillkommene Müller Charlotten und dem Hauptmann zugebracht, unter **den Bäumen** ein architektonischer Ruhesitz aufgeführt werden, wonach diejenigen, die über den See fahren, zu steuern hätten. "wo wird man denn nun abendlich umgibt. mit **den Bäumen**, die um uns blühen, grünen, Frucht tragen, mit jeder Staude, an der wir vorbeigehen, mit jedem Grashalm, über den wir hinwandeln, habe auszubreiten, wenn **die Bäume** so geisterhaft, so durchsichtig vor uns stehen. sie sind nichts, aber sie denken auch nichts zu. wie aber einmal Knospen und Blüten kommen reifen Früchte an **den Bäumen**, die man damals blühen gesehen. der Major ging ab und zu; auch Mittler ließ sich öfter sehen. die Abendstimmungen waren meistens regeln

About KorAP? V 0.18.0

Abb. 3: Die KorAP-Web-Oberfläche erlaubt den direkten Zugriff auf die Annotationsdaten aller Foundries

Um die Handhabbarkeit von Annotationen zu verbessern, bietet KorAP die Möglichkeit, prinzipiell beliebig viele konkurrierende Annotationen zu verwenden und abzufragen. So können z.B. – je nach Aufgabenstellung – zur Maximierung des Recalls bzw. zur Minimierung von falsch Negativen disjunktive (Tab. 1, Nr. 3) und zur Maximierung der Precision (bzw. zur Minimierung von falsch Positiven) konjunktive Anfragen (Tab. 1, Nr. 4) über verschiedene Annotationsschichten verwendet werden. Wie Abbildung 4 und Tabelle 1 zeigen, können selbst bei einer relativen hohen Übereinstimmung zwischen zwei Taggern die so erzielten Trefferraten deutlich voneinander abweichen.

Außerdem können je nach zu untersuchendem Phänomen und nach zu untersuchender Sprachdomäne gezielt die Annotationen verwendet werden, die die besten Ergebnisse versprechen. Dies kann umso sinnvoller sein, wenn die Annotationen nicht nur auf unterschiedlichen Algorithmen, sondern auch auf unterschiedlichen Trainingsdaten beruhen.

Tab. 1: Suche nach ‚das‘, annotiert als Relativpronomen von: (1) TreeTagger, (2) CoreNLP, (3) TreeTagger oder CoreNLP, (4) TreeTagger und CoreNLP mit der Anzahl der jeweiligen Treffer in DeReKo-2016-I

Nr.	Anfrage (Poliquarp-Syntax)	Mio. Treffer
1	[orth="das" & tt/p=PRELS]	6,75
2	[orth="das" & corenlp/p=PRELS]	6,49
3	[orth="das" & (tt/p=PRELS corenlp/p=PRELS)]	7,60
4	[orth="das" & (tt/p=PRELS & corenlp/p=PRELS)]	5,65

Abb. 4: Konfusionsdiagramm zur Annotation der Zeichenkette ‚das‘ als Artikel (ART), substituierendes Demonstrativpronomen (PDS) oder substituierendes Relativpronomen (PRELS) in DeReKo: TreeTagger vs. CoreNLP (german-fast.tagger). Die Übereinstimmung der beiden Tagger liegt bzgl. ‚das‘ bei ca. 91% ($\kappa=0.77$)

4 Virtuelle Korpora

Eine weitere Verbesserung gegenüber COSMAS II betrifft die Konstruierbarkeit virtueller Korpora. Diese können in KorAP dynamisch und persistent für die weitere Verwendung anhand von Metadatenvariablen wie Texttyp, Thema, Publikationsort oder -zeit vom Nutzer definiert werden. Zum einen ist dies gemäß dem DEREKo-Design-Konzept der *Ur-Stichprobe* (Kupietz/Keibel 2009) essenziell, da so der Nutzer in die Lage versetzt wird, selbst stratifizierte Substichproben aus DEREKo zu ziehen, die die bzgl. der jeweiligen Fragestellung relevante Sprachdomäne möglichst gut abbilden (vgl. Kupietz 2015). Speziell relevant für die Variationsforschung ist zudem ihre Verwendung im Rahmen kontrastiver Untersuchungen etwa zu Varietäten oder zu anderweitig definierten Sprachdomänen, wenn diese im Hinblick auf mehrere Fragestellungen untersucht werden sollen.

5 Quantitative Funktionalitäten

Im Bereich quantitativer Analysen, ihrer Darstellung, Interpretation und Auswertung sind in KorAP gegenüber COSMAS II einige Verbesserungen geplant. So sollen z.B. Trefferfrequenzen in Abhängigkeit von Metadatenvariablen abgefragt, dargestellt, statistisch ausgewertet und exportiert werden können. Diese Funktionalitäten sind ebenso wie die Kookkurrenzanalyse zurzeit allerdings noch nicht implementiert bzw. nicht integriert, da die zugrundeliegenden Frequenzfunktionen z.T. kritisch von der noch nicht abgeschlossenen Implementation von *Kanalito*, der KorAP-Komponente die die Suchresultate aggregiert, abhängig sind. Um auch komplexe Frequenzanfragen möglichst bald zu unterstützen, wird nach der Fertigstellung von Kanalito das Hauptaugenmerk auf der Entwicklung der API liegen. Über KorAPs REST-API, die auch von KorAPs Web-Oberfläche *Kalamar* verwendet wird, können dann beliebige Clients, z.B. R-Skripte oder andere Web-Applikationen, den gesamten Funktionsumfang ausnutzen und die JSON-kodierten Resultate weiterverarbeiten, auswerten und visualisieren.

Bei KorAPs eigenen Auswertungs- und Visualisierungsfunktionen wird ein Hauptaugenmerk darauf liegen, eine enge Verzahnung von quantitativen Analysen und qualitativen Interpretationen möglichst gut zu unterstützen. Visualisierungen von Analyseergebnissen sollen z.B. in der Weise interaktiv gestaltet werden, dass sie zum einen eine möglichst schnelle Exploration großer Treffermengen und die Abduktion von Hypothesen ermöglichen. Zum anderen sollen die aggregierten Darstellungskomponenten falls möglich immer mit ihrer Provenienz

nienz, also den durch sie zusammengefassten Korpusdaten, verlinkt sein, um Hypothesen möglichst schnell anhand der Primärdaten überprüfbar zu machen (siehe Abb. 5). Eine solche Verlinkung ist auch durch externe Applikationen leicht realisierbar, da alle KorAP-Anfragen in Form von URLs kodiert werden können.

Abb. 5: Kontrastive Darstellung von Frequenzverläufen mit Konfidenzintervallen und Verlinkung der einzelnen Datenpunkte zu den aggregierten Suchtreffern (bzw. den entsprechenden Suchanfragen)

6 Nachhaltigkeit und Erweiterbarkeit

Neben den bisher genannten und einigen weiteren Funktionalitäten ist eines der Hauptziele von KorAP, eine Nachhaltigkeit und Wartbarkeit der Kernkomponenten von 15–20 Jahren nach öffentlicher Freischaltung zu erreichen. Eine wichtige Design-Entscheidung, um dieses Ziel trotz wachsender Anwendungs- und Methodenspektren in der Korpuslinguistik zu erreichen, war, ein erweiterbares System aus leicht ersetzbaren Komponenten zu entwickeln (i.d.R. so genannte Microservices; Diewald et al. 2016), statt ein monolithisches System zu konzipieren, das von Beginn an alle heutigen und zukünftigen Anforderungen erfüllen kann. Die Kernkomponenten sollen dabei priorisiert entwickelt und intern verwendete APIs, soweit rechtlich möglich, auch externen Applikationen zur Verfügung

gestellt werden. Darüber hinaus laden wir alle Nutzer, die zusätzliche Kernfunktionalitäten benötigen, dazu ein, selbst an der Weiterentwicklung der Open-Source-Software KorAP mitzuarbeiten.³

7 Fazit

KorAP wird durch die Unterstützung prinzipiell unbegrenzt großer Korpora mit beliebigen und beliebig vielen Annotationsschichten, mehrerer erweiterbarer Anfragesprachen und virtueller Korpora sowie durch die geplanten Funktionalitäten zur quantitativen Auswertung die Möglichkeiten zur Erforschung grammatischer Variation verbessern. Darüber hinaus ist der Funktionsumfang durch externe Applikationen, die auf die öffentliche API zugreifen können, und durch die Mitarbeit am Open-Source-Projekt KorAP erweiterbar.

Literatur

- Bański, Piotr/Frick, Elena/Witt, Andreas (2016): Corpus Query Lingua Franca (CQLF). In: Calzolari (Hg.), S. 2804–2809.
- Belica, Cyril et al. (2011): The morphosyntactic annotation of DeReKo: Interpretation, opportunities, and pitfalls. In: Konopka, Marek et al. (Hg.): Grammatik und Korpora 2009. Dritte internationale Konferenz, Mannheim, 22.–24. September 2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen, S. 451–469.
- Bingel, Joachim (2015): Instantiation and implementation of a Corpus Query Lingua Franca. Masterarbeit, Universität Heidelberg.
- Bingel, Joachim/Diewald, Nils (2015): KoralQuery – a general corpus query protocol. In: Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, Vilnius, 11. – 13. Mai 2015. (= NEALT Proceedings Series). Linköping. Internet: www.ep.liu.se/ecp/111/001/ecp15111001.pdf (Stand: 22.8.2016).
- Bodmer, Franck (1996): Aspekte der Abfragekomponente von COSMAS II. In: LDV-INFO 8, S. 142–155.
- Calzolari, Nicoletta et al. (Hg.) (2016): LREC 2016 – Proceedings of the Tenth International Conference on Language Resources and Evaluation, Portorož, 23.–28. Mai 2016. Paris.
- Diewald, Nils et al. (2016): KorAP Architecture – Diving in the deep sea of corpus data. In: Calzolari (Hg.), S. 3586–3591.

³ Der KorAP-Quell-Code ist unter der BSD-2-Lizenz auf <https://github.com/KorAP> und <https://korap.ids-mannheim.de/gerrit> veröffentlicht.

- Giesbrecht, Eugenie/Evert, Stefan (2009): Part-of-Speech (POS) tagging – a solved task? An evaluation of POS taggers for the Web as Corpus. In: Alegria, Inaki/Leturia, Igor/Sharoff, Serge (Hg.): Proceedings of the Fifth Web as Corpus Workshop (WAC5). Pre-SEPLN Workshop, San Sebastian, 7. September 2009. San Sebastian, S. 27–36.
- Kupietz, Marc (2015): Constructing a corpus. In: Durkin, Philip (Hg.): The Oxford handbook of lexicography. Oxford, S. 62–75.
- 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 3. Tokyo, S. 53–59.
- Kupietz, Marc/Lüngen, Harald (2014): Recent developments in DeReKo. In: Calzolari, Nicoletta et al. (Hg.): LREC 2014 – Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, 26. – 31. Mai 2014. Paris S. 2378–2385.
- Kupietz, Marc et al. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (Hg.): LREC 2010 – Proceedings of the Seventh Conference on International Language Resources and Evaluation, 17.–23. Mai 2010, Malta. Paris, S. 1848–1854.
- Meurers, Detmar/Müller, Stefan (2009): Corpora and syntax. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus linguistics. 2. Halbbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.2). Berlin, S. 922–933.
- Rosenfeld, Viktor (2010): An implementation of the Annis 2 query language. Technical report, Humboldt-Universität zu Berlin. Berlin.
- Sporny, Manu et al. (2014): JSON-LD 1.0. A JSON-based serialization for linked data. Technical report, W3C. W3C Recommendation, 16. Januar 2014. Internet: www.w3.org/TR/json-ld/ (Stand: 10.8.2016).