Leibniz-Gemeinschaft

Abschließender Sachbericht

Titel des Vorhabens:

Korpusanalyseplattform der nächsten Generation

Leibniz-Einrichtung: Institut für Deutsche Sprache
Aktenzeichen: SAW-2011-IDS-2
Projektlaufzeit: 01.07.2011 – 30.06.2015
Ansprechpartner: Dr. Marc Kupietz <kupietz@ids-mannheim.de>

Inhaltsverzeichnis

1. Executive Summary

Systematisch zusammengestellte Sammlungen von Texten mit mehreren Milliarden Wörtern, so genannte *very large corpora*, dienen in zunehmendem Maße als die hauptsächliche empirische Grundlage für fast alle Bereiche der Linguistik. Sie werden zum einen dazu verwendet, bestehende Hypothesen zum Sprachgebrauch zu überprüfen und zum anderen dazu, zu neuen Hypothesen über Strukturen, Gesetzmäßigkeiten, Regularitäten und Funktionen von Sprache zu gelangen. Um Sprachforscher in die Lage zu versetzen, solche Untersuchungen auf ausreichend großen Datenmengen durchzuführen, sind geeignete Software-Werkzeuge unabdingbar.

Mit der *Korpusanalyseplattform der nächsten Generation (KorAP)* wurde im Rahmen dieses Leibniz-Wettbewerb-Vorhabens am Institut für Deutsche Sprache in der Tradition von REFER (1983), COSMAS (1994) und COSMAS II (2003) ein neues wissenschaftliches Werkzeug entwickelt, das diesen Bedarf decken soll und insbesondere den derzeit etwa 39.000 COSMAS-II-Nutzern aus der germanistischen Linguistik ein nachhaltiges wissenschaftliches Werkzeug für den methodisch validen Umgang mit *very large corpora* an die Hand gibt.

Die Herangehensweise bei der Entwicklung von KorAP war, von vornherein kein unrealistisches System anzustreben, das alle Features implementiert, die in den nächsten 15 Jahren in der Linguistik und den angrenzenden Disziplinen benötigt werden, sondern stattdessen eine Plattform zu entwickeln, die zwar die Funktionalitäten bisheriger Systeme übertrifft, aber insbesondere alle Grundvoraussetzungen erfüllt, um eine stetige Erweiterbarkeit über diesen Zeitraum auch für spezialisierte Anwendungen für externe Benutzer bei einem realistischen Wartungsaufwand zu ermöglichen. Zu diesen Grundvoraussetzungen gehörte insbesondere die Unterstützung prinzipiell unbeschränkter Mengen an Primär- und Annotationsdaten, die Unterstützung einer unbeschränkten Anzahl beliebiger linguistischer Annotationsschichten, Definierbarkeit von stratifizierten Sub-Stichproben, Schnittstellen zur Anbindung externer Systeme und neuer Entwicklungen sowie ein Rechtemanagement, das der besonderen Abhängigkeit linguistischer Forschungsprimärdaten von Rechten Dritter Rechnung trägt. Als Basis für die Erfüllung dieser Voraussetzungen wurde für KorAP eine horizontal skalierbare Micro-Service-Architektur gewählt, deren auf leichte Austauschbarkeit ausgerichtete Komponenten möglichst auf Open-Source-Bibliotheken mit breiter Entwicklerbasis aufsetzen, und deren Schnittstellen – soweit sinnvoll und rechtlich möglich – auch von zukünftigen und externen Komponenten genutzt werden können. In diesem Kontext neue Ansätze wurden dabei z. B. für die möglichst weitgehende Ausschöpfung von Lizenzrechten durch die Übertragung von Jim Greys Prinzip *put the computation near the data*, die sparsame intensionale Repräsentation von virtuellen Korpora anhand von Suchanfragen, die effiziente, informationell-abgeschlossene und transparente Zugriffskontrolle durch das Umschreiben von Suchanfragen (*query rewriting*) sowie die Unterstützung einer erweiterbaren Menge von Anfragesprachen mittels eines internen Anfrageprotokolls gewählt.

Große Teile von KorAP wurden bereits unter der vereinfachten BSD-Lizenz open-source veröffentlicht. Weitere werden folgen. Am IDS wird KorAP durch das Dauerprojekt Korpusrecherchesystem weiterentwickelt. Im Laufe der nächsten 2,5 Jahre wird es schrittweise sein Vorgängersystem COSMAS II ablösen.

2. Ausgangsfragen und Zielsetzung des Vorhabens

2.1. Ausgangsfragen

Am Institut für Deutsche Sprache sind bereits seit Ende der 60er Jahre die größten Korpora des geschriebenen Deutsch beheimatet (Teubert / Belica 2014). Diese dienen der Bestätigung oder Widerlegung von Hypothesen und als Grundlage explorativer Forschungsarbeiten in der germanistischen Linguistik weltweit. Um große Korpora für Sprachwissenschaftler handhabbar zu machen, sind außerdem geeignete Werkzeuge unabdingbar, die sehr große Datenmengen unterschiedlicher Modaliät persistent verwalten und methodisch valide analysieren können. Mit dem System zur Korpusanalyse und -recherche COSMAS II (Bodmer 1996) sowie einer größeren Anzahl ähnlicher Anwendungen verfügte das IDS bereits zu Beginn des Vorhabens über ein breites Spektrum an Handwerkszeugen zur Auswertung solcher Textsammlungen. In der Linguistik und angrenzenden Disziplinen gab es in den letzten Jahren jedoch Entwicklungen, die Anpassungen an die bisher angewandten Methoden und Werkzeuge notwendig machten. Mit der zunehmenden Empirisierung der Geisteswissenschaften und des zunehmenden Stellenwerts von Korpora als empirische Hauptgrundlage in der Linguistik wurden Korpusanalysesysteme für ein immer breiter werdendes Spektrum an Nutzern und Nutzungsarten relevant. Zudem war ein verstärkter Trend zu kollaborativer Arbeit in verteilten Rechnernetzen auch in der Linguistik zu beobachten. Dies bedeutete für neue Entwicklungen, dass Schnittstellen für solche Infrastrukturen geboten werden müssen, die nachnutzbare Such- und Analyse-Schemata sowie die Rückeinspeisung von Benutzerannotationen ermöglichen. Neue Anforderungen ergaben sich zudem aus dem immensen Wachstum von Korpora. Dies betraf zum einen die Notwendigkeit, Textmengen im Umfang von mehreren 10 Milliarden Wörtern einschließlich multipler Annotationen entsprechend linguistischer Anforderungen, die weit über die des reinen Information Retrieval hinausgehen, zu indizieren und nutzbar zu machen. Zum anderen ergaben sich aus dem Korpuswachstum neue qualitative Anforderungen, da auf der Grundlage großer Stichproben über rein lexikalische Fragestellungen hinaus komplexere sprachliche Muster und Strukturen analysiert werden können.

2.2. Zielsetzung

Im einzelnen ergaben sich aus den oben genannten allgemeinen Voraussetzungen und Entwicklungen sowie den Erfahrungen mit dem Vorgänger-System und seinen Nutzern die folgenden Zieleigenschaften für KorAP:

1. Tragfähigkeit des Basissystems für mindestens 12-15 Jahre
2. Erweiterbarkeit – insbesondere auch durch extern entwickelte Komponenten
3. methodische Validität
4. prinzipielle Unbeschränktheit der Menge an Primär- und Annotationsdaten
5. prinzipielle Unbeschränktheit der Menge an Annotationsschichten
6. Unterstützung von Annotationen, insbesondere zu:
    a. Morphosyntax
    b. Konstituenz
    c. Dependenz
7. Unterstützung der wesentlichen Funktionalitäten des Vorgängersystems COSMAS II
8. leichte Erweiterbarkeit von Analyse- und Visualisierungsfunktionalitäten

9. gegenüber COSMAS II verbesserte Unterstützung von einerseits gelegentlichen Nutzern und andererseits Experten-Nutzern
10. umfangreiche Möglichkeiten zur stratifizierten Ziehung von Sub-Korpora (virtuelle Kollektionen / Korpora) anhand von Meta- und Primärdaten durch den Nutzer (Kupietz et al. 2010)
11. persistente Referenzierbarkeit und Rekonstruierbarkeit von virtuellen Korpora / Kollektionen (entsprechend ISO 24619, Broeder et al. 2007)
12. exakte Abbildung von Lizenzbedingungen bzgl. der von Rechten Dritter betroffenen Textdaten im Rechtemanagement, abhängig von
    a. Nutzer / Nutzergruppe
    b. Lizenztyp der Ressource
13. möglichst weitgehende Ausschöpfung von Nutzungsrechten an Korpustexten durch den Endnutzer
14. Anschlussfähigkeit an externe Infrastrukturen wie CLARIN[1]
15. möglichst weitgehende Verwendung von Standard-Formaten, -Schnittstellen und -Protokollen

3. Entwicklung der durchgeführten Arbeiten

Die Arbeiten begannen zunächst mit Vorstudien zu den Bereichen:

1. generelle Systemarchitektur
2. Datenmodellierung:
    a. Standoff-Annotationen
    b. Annotationen und ihre Gruppierung in *Foundries*
    c. Darstellung von Relationen
    d. Metadaten
3. mögliche DBMS-Grundlage(n):
    a. RDBMS (Ergebnisse in Schneider 2012)
    b. basale Key / Value-Datenbank wie Berkeley-DB
    c. BigTable-DB
    d. Volltextdatenbanken
    e. Graphdatenbanken
    f. XML-Datenbanken
4. Anfragesprachen
5. Frontend-Framework
6. Ermittlung von möglicherweise noch unbekannten Nutzerwünschen durch Interviews mit den Abteilungen Lexik, Grammatik und Pragmatik des IDS, jeweils stellvertretend für die jeweiligen Forschungs- und Anwendungskontexte

[1] http://clarin.eu/

Nach der Entscheidung für eine microservice-basierte Architektur, Lucene als Grundlage für die Hauptdatenbankkomponente, Vaadin als vorläufiges Framework für die Benutzeroberfläche, JAVA als hauptsächliche Entwicklungssprache, Poliqarp (Przepiórkowski et al. 2004) als Grundlage für eine Hauptanfragesprache und der Fertigstellung eines Book of Use-Cases, wurde die Arbeit nach etwa 7 Monaten Projektlaufzeit in den folgenden Arbeitspaketen organisiert fortgesetzt:

1. Datenbank-Backend und Datenrepräsentation
    a. Lucene-basiert
    b. Neo4J-basiert (ab Mitte 2013)
2. Frontend
3. Anfragesprachen (und Standardisierung)
4. Anfrageprotokoll (ab Mitte 2013)
5. Benutzer- und Rechtemanagement
6. Korpusimport
    a. verschiedene Korpusenkodierungsformate, insbesondere I5 (TEI-Customisierung für DeReKo; Lüngen / Sperberg Mc-Queen, 2012)
    b. Anwendung und Integration von Annotationswerkzeugen

Die Arbeiten an 1a, 2, 3 und 4 konnten im letzten Projektjahr für die Veröffentlichung der Source-Codes in erster Version abgeschlossen werden.[2] Die open-source Veröffentlichung von 6 ist problematisch, da dort externe Annotationstools mit nicht-kompatiblen, zum Teil kommerziellen, Lizenzen einbezogen werden, die zum Teil keine Veröffentlichung erlauben. Als Kompromisslösung ist geplant, die Korpusimport-Komponente in einer Version zu veröffentlichen, die nur ausreichend und kompatibel lizenzierte Komponenten enthält.

Außerdem stand zum Ende der Projektlaufzeit noch die Open-Source-Veröffentlichung des Benutzer- und Rechtemanagements (5) aus. Diese Komponente konnte erst nach Ende der Projektlaufzeit in erster Version fertiggestellt werden und soll aufgrund ihrer rechtlichen Relevanz vor ihrer Veröffentlichung im beta-Produktionsbetrieb getestet werden.

3.1. Abweichungen vom ursprünglichen Konzept

Die Arbeiten konnten im Großen und Ganzen wie geplant durchgeführt werden. Es ergaben sich jedoch einige Verzögerungen, so dass die ursprünglich auf 3 Jahre geplante Projektlaufzeit zuwendungsneutral um ein Jahr verlängert werden musste. Die Gründe dafür waren, dass der KorAP-Hauptentwickler das IDS nach zwei Jahren Projektlaufzeit verließ und zwei Mitarbeiter zeitgleich in Elternzeit gingen. Zudem konnten die Stellen aufgrund der hohen technischen Anforderungen kombiniert mit sprachwissenschaftlichen Anforderungen nicht kurzfristig und für die kurze restliche Projektlaufzeit wiederbesetzt werden.

Kleinere Abweichungen vom Arbeitsplan ergaben sich zum einen dadurch, dass bereits während der Projektlaufzeit die Menge der Korpus-Primärdaten um den Faktor 4 wuchs (Kupietz et al. 2014). Um dieses Wachstum abzufedern und da gleichzeitig ein Projekt zur Langzeitarchivierung am IDS entstand, wurde entschieden, dass die Verwaltung von Versionsgeschichten der Korpusdaten nicht Aufgabe des KorAP-Systems, sondern Aufgabe des Systems für die Langzeitarchivierung sein soll.

[2] http://github.com/KorAP/

Zum anderen wuchs während der Projektlaufzeit auch bei anderen nationalen Sprachinstituten (Polen, Ungarn, Rumänien, Österreich, Tschechien) und seitens des Rats für deutsche Rechtschreibung das Interesse an einem System wie KorAP, so dass die Prioritäten leicht in Richtung einer langfristigen kollaborativen Weiterentwicklung von KorAP, einer standortverteilten Datenhaltung und der Möglichkeit kontrastiver Untersuchungen auf vergleichbaren Korpora verschoben wurden.

4. Darstellung der erreichten Ergebnisse

Die oben genannten Hauptziele konnten im Wesentlichen erreicht werden. KorAP wurde in einer ersten alpha-Version mit begrenzter Funktionalität am 17.02.2014 im IDS vorgestellt. Seitdem ist eine aktuelle alpha- bzw. beta-Version IDS-intern für Testzwecke zugänglich. Die öffentliche Zugänglichmachung für die 39.000 Nutzer des Vorgängersystems COSMAS II und andere Interessierte ist für das Frühjahr 2016 geplant. Die Wartung und Weiterentwicklung von KorAP hat nach Ablauf der SAW-Förderung zum 01.07.2015 das neue IDS-Dauerprojekt *Korpusrecherchesystem* übernommen, das z. T. aus dem COSMAS-II-Projekt hervorgegangen ist.

Im Folgenden werden die einzelnen Komponenten und Funktionalitäten des KorAP-Systems hinsichtlich der oben genannten Zielsetzungen mit den dazu gewählten Lösungsstrategien dargestellt.

4.1. Gesamtsystem

Um die Anforderungen hinsichtlich der Nachhaltigkeit des Systems zu erfüllen, wurde eine microservice-basierte Architektur für KorAP entwickelt, in der einzelne Komponenten erweiterbar, austauschbar und leicht wartbar gehalten werden. Um eine horizontale Skalierbarkeit zu ermöglichen, kann das Datenbank-Backend auf beliebig viele Knoten verteilt werden (s. Abb. 1).

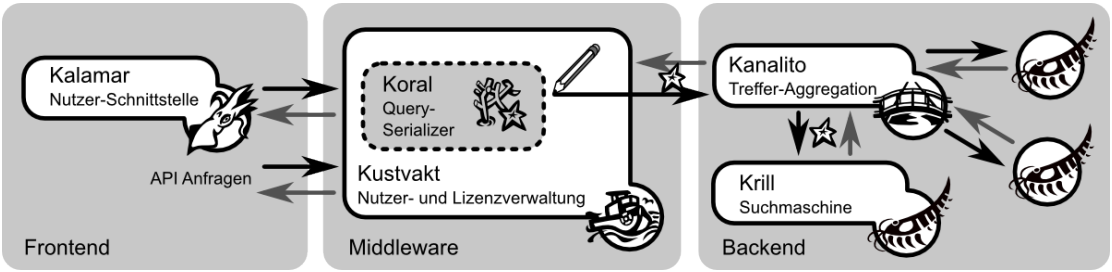

Abbildung 1: Gesamtarchitektur des KorAP-Systems

4.2. Datenbankkomponenten

Als Ergebnis der Vorstudien zu möglichen Datenbanksystemen wurde für die Such- und Analysekomponente mit Apache Lucene[3] eine etablierte, freie Suchtechnologie gewählt, nicht zuletzt um auch einen möglichst hohen Nutzen aus Fremdentwicklungen zu schöpfen. Insbesondere die hohe Performanz und flexible Einsatzmöglichkeit von Lucene war für diese Entscheidung ausschlaggebend.[4] Die darauf aufbauende Komponente *Krill* unterstützt eine große Zahl ver-

[3] https://lucene.apache.org/

[4] Lucene wird inzwischen auch für andere Korpussuchmaschinen verwendet, wie z.B. für das am Instituut voor Nederlandse Lexicologie (INL) sich in Entwicklung befindliche BlackLab (s. https://github.com/INL/BlackLab/ und Evert / Hardie 2015) und eine ähnliche Entwicklung am Meertens Instituut (Brouwer et al. i. V.).

schiedener Suchoperationen auf den Primär- und Annotationsdaten, wie sie in den unterstützen Anfragesprachen definiert sind. Zudem lassen sich virtuelle Kollektionen auf Basis von Metadaten definieren. Hierfür wurden die internen Query-Funktionalitäten von Lucene erheblich erweitert und die Indexstrukturen zur Speicherung von Annotationen angepasst.

Für die Verteilung wurden etablierte, auf Lucene aufsetzende Technologien wie Solr[5] und ElasticSearch[6] evaluiert, letztendlich aber eine einfache, eigene Komponente, *Kanalito*, entwickelt, die Suchergebnisse mittels einer PostgreSQL-Datenbank[7] aggregiert.

Entsprechend dem Konzept, für verschiedene Arten von Anfragen unterschiedliche, spezialisierte Datenbank-Indizes zu ermöglichen und als Testfall für die generelle Austauschbarkeit der Datenbankkomponente, wurde speziell für komplexe Relationsanfragen auf Dependenz-Annotationen ein weiteres Datenbank-Backend entwickelt, das auf der Graph-Datenbank Neo4j[8] beruht. Dabei konnte auf Vorarbeiten im Zusammenhang mit dem Polnischen Nationalkorpus aufgebaut werden (s. Pęzik 2013). Das Neo4j-Backend befindet sich z. Zt. noch im Alpha-Stadium und wird zunächst nicht produktiv eingesetzt.

4.3. Nutzerverwaltung und Rechtemanagement

Die Sprachwissenschaft hat anders als die meisten anderen Disziplinen das Grundproblem, dass ihre Forschungsprimärdaten, also Texte und aufgezeichnete Sprache, von Rechten Dritter (Urheberrecht, Datenbankrechte, allgemeine Persönlichkeitsrechte, usw.) betroffen sind (vgl. Kupietz 2015: 73). Da diese Rechteinhaber, wie Verlage, Autoren und Sprecher, außerdem nicht Teil der wissenschaftlichen Community sein müssen, können Open-Access-Praktiken aus anderen Disziplinen nicht auf die Sprachwissenschaft übertragen werden. Um Texte für die eigene Forschung und die Nutzung durch andere Wissenschaftler verfügbar zu machen, müssen daher in der Regel individuelle Lizenzverträge mit Rechteinhabern abgeschlossen werden, bei denen typischerweise die eingeräumten Rechte in direktem Zusammenhang mit den verlangten Lizenzgebühren stehen.

Aufgrund dieser Problematik ist das Rechtemanagement für ein System wie KorAP, das Linguisten die Forschung auf großen Textmengen sehr vieler unterschiedlicher Rechteinhaber erlaubt, von hoher rechtlicher, wirtschaftlicher und wissenschaftlicher Relevanz. Die kritischen Anforderungen sind dabei insbesondere 1) Nutzungsrechte abhängig vom Text und vom Nutzer möglichst exakt abzubilden, um einerseits alle vorhandenen Rechte ausschöpfen zu können und andererseits keine Lizenzbedingungen zu verletzten, 2) trotz der komplexen Rechtemodellierung einen ausreichend effizienten Zugriff zu ermöglichen und 3) den Benutzer möglichst darauf hinzuweisen, wenn auf erwartete Daten aufgrund mangelnder Rechte nicht zugegriffen werden kann.

Um diese Anforderungen gleichermaßen zu erfüllen, wurde eine Strategie gewählt, die darauf beruht, standardisiert serialisierte Anfragen für den Nutzer transparent umzuschreiben (Bański et al. 2014).

Das Rechtemanagementsystem basiert auf spezialisierten Anfragen an eine relationale Datenbank. Gemäß den Anforderungsprofilen der Dokumentsammlungen, auferlegt durch Lizenzvereinbarungen mit Rechteinhabern, wurden standardisierte Programmschnittstellen entwickelt, durch die sich zugriffsrelevante Informationen extrahieren und verarbeiten lassen. So können zum einen spezifische Dokumentensammlungen gefiltert, sowie weitere Anfrageparameter angepasst werden, falls sie nicht den hinterlegten Richtlinien entsprechen (Bański et al. 2013). Zum

[5] https://lucene.apache.org/solr/
[6] https://www.elastic.co/products/elasticsearch
[7] http://www.postgresql.org/
[8] http://neo4j.com/

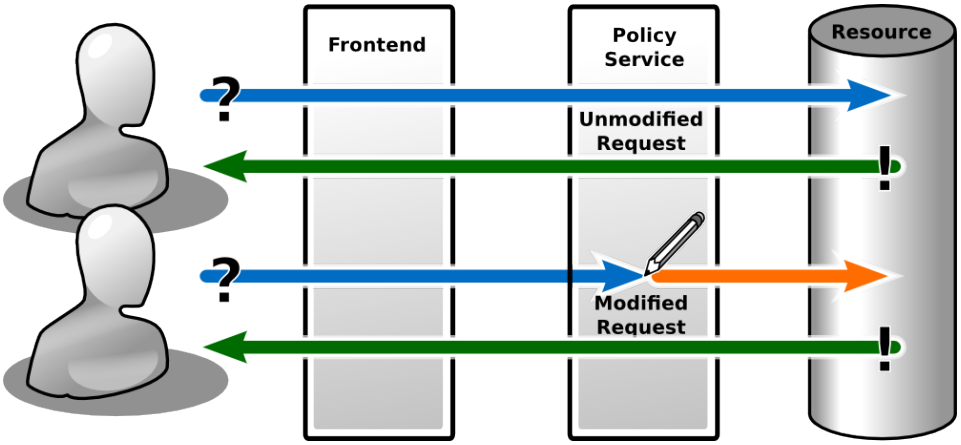

Abbildung 2: *Query-Rewriting:* Falls Anfragen an bestimmte Ressourcen nicht berechtigt sind, werden sie durch den Policy-Service modifiziert.

anderen können über eine in KoralQuery implementierte Nachrichtenfunktion dem Benutzer Meldungen über etwaige Änderungen bzw. Zugriffsbeschränkungen der Suchanfrage zurückgeliefert werden.

Zur Authentifizierung unterstützt das Rechtemanagement das Shibboleth-Framework und die Protokolle OAuth2.0 und OpenID Connect, über die sich nicht nur Nutzer mittels Browser, sondern auch externe Applikationen und Services authentifizieren können, um die KorAP-API zu verwenden.

4.4. Anfragesprachen

Bzgl. der Suchanfragesprachen bestanden folgende Hauptanforderungen:

1. die Anfragesprache des Vorgängersystems COSMAS II mit seinen 39.000 Nutzern (hauptsächlich aus der germanistischen Linguistik) sollte weiterhin unterstützt werden
2. in einer Suchanfrage sollten Annotationen unterschiedlicher Quellen referenzierbar sein
3. die anhand von Vorstudien avisierte Mächtigkeit sollte weitgehend abgedeckt sein (s. Frick et al. 2012, Kupietz / Frick 2012)
4. außerdem sollten möglichst durch die Unterstützung weiterer Anfragesprachen weitere Nutzerkreise erschlossen werden.

Um die Anforderungen zu erfüllen, wurden die zu unterstützenden Suchoperatoren der Zielsprachen in einer allgemeinen Protokollsprache, *KoralQuery*, zusammengeführt (Bingel / Diewald 2015; Diewald / Bingel 2015). Hierbei werden identische Operatoren unterschiedlicher Anfragesprachen identisch serialisiert. KoralQuery basiert auf JSON-LD (Sporny et al. 2014) und verfolgt, anders als die unterstützten Anfragesprachen, nicht das Ziel einfacher Formulierbarkeit, sondern möglichst uneingeschränkter Aussagemächtigkeit durch leichte Erweiterbarkeit bei guter Kompatibilität zwischen allen Komponenten der KorAP Architektur.

Im Projektzeitraum konnte die Ausdrucksmächtigkeit der Anfragesprachen von Cosmas-II (Bodmer 1996; Stand zum Projektbeginn), Poliqarp (Przepiórkowski et al. 2004) und Annis QL (Rosenfeld 2010) abgebildet werden. Zudem wird jene Teilmenge von CQL unterstützt, die für die Einbindung von KorAP in die CLARIN Federated Content Search Infrastruktur[9] erforderlich

[9] https://www.clarin.eu/content/federated-content-search-clarin-fcs

ist. Neu entwickelte Konzepte, die nicht Teil der unterstützten Sprachen sind, wurden in eine Erweiterung der Poliqarp-Anfragesprache (Poliqarp+) integriert.

Alle Anfragen werden durch ein zentrales Transformationsmodul, *Koral* (Bingel 2015), in KoralQuery übersetzt. Koral nutzt das ANTLR-Framework (Parr / Quong 1995), um verschiedene Anfragen über definierte Grammatiken in abstrakte Syntaxbäume zu übertragen und diese als KoralQuery zu serialisieren. Damit ist Koral leicht durch weitere Grammatiken erweiterbar.

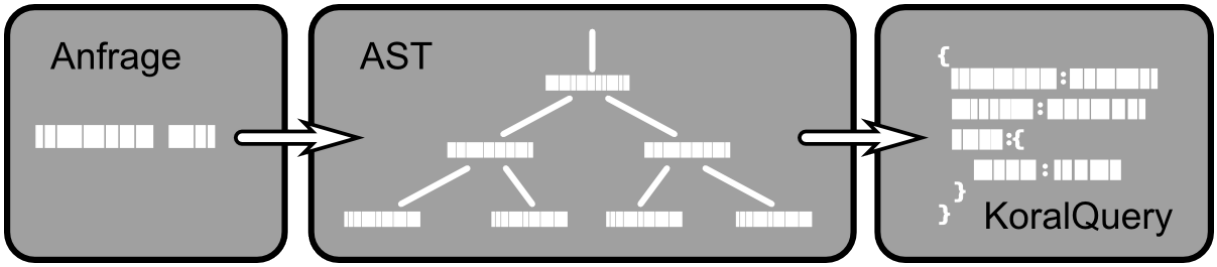

Abbildung 3: Koral interpretiert alle Eingaben verschiedener Anfragesprachen zunächst als abstrakten Syntaxbaum (AST) und wandelt diesen schließlich in KoralQuery um.

Es ist anzumerken, dass die Ausdrucksmächtigkeit von KoralQuery nicht zwingend von allen KorAP Komponenten unterstützt wird und unterstützt werden muss: Die Protokollsprache wurde konzipiert, um eine möglichst große Kompatibilität zu bieten und bei nicht unterstützten Suchoperatoren erwartbar zu reagieren.

Im Zuge der Vorarbeiten zur Festlegung der KorAP-Anfragesprache wurde, u. a. um die Ausdrucksmächtigkeit von Korpusanfragesparchen formalisieren und vergleichen zu können, ein Entwurf für eine Corpus Query Lingua Franca (CQLF) entwickelt und in den Standardisierungsprozess der zuständigen ISO-Arbeitsgruppe ISO/TC 37/SC 4/WG 6 eingebracht. Dieser wurde als neuer ISO-Arbeitsgegenstand akzeptiert und hat z. Zt. im ISO-Standardisierungsprozess den Status eines *Committee Draft* (ISO/CD 24623-1). Aus der Entwicklung von CQLF ging u. a. auch die Idee zu einem allgemeinen internen Anfrageprotokoll hervor.

4.5. Frontends

Die grafischen Nutzeroberflächen setzen auf der API des KorAP-Systems auf und sind daher leicht austauschbar und an spezifische Anforderungen verschiedener Projekte leicht anpassbar. In der Projektlaufzeit von KorAP wurden zwei grafische Benutzeroberflächen entwickelt: Zum einen ein Frontend auf Basis des Java-Web-Frameworks Vaadin[10], das zu Beginn entwickelt wurde, um ein möglichst homogenes Entwicklungsumfeld zu bieten (in welchem alle Komponenten in der selben Programmiersprache geschrieben wurden). Zum anderen ein Frontend auf Basis des Perl-Web-Frameworks Mojolicious[11], das größere Flexibilität in Bezug auf die Einbindung clientseitiger Skripte in JavaScript bietet. Aufgrund des inzwischen größeren Funktionsumfangs und des geringeren Entwicklungsaufwands ist die Mojolicious-basierte grafische Nutzerschnittstelle *Kalamar* derzeit das Primär-Frontend von KorAP (siehe Abb. 4).

Bei der Entwicklung von Kalamar wurde ein besonderer Fokus auf einen möglichst einfachen Zugriff auf Primär- und Annotationsdaten gelegt. Hierbei werden je nach Datenform unterschiedliche Darstellungen eingesetzt (KWIC-, Tabellen- oder Baumansicht). Visuelle Hilfen zur Suche in verschiedenen Annotationen und zur Erstellung virtueller Kollektionen sowie ein eingebettetes Handbuch unterstützen Nutzer bei der Suche in großen Datenmengen.

[10] https://vaadin.com/
[11] http://mojolicio.us/

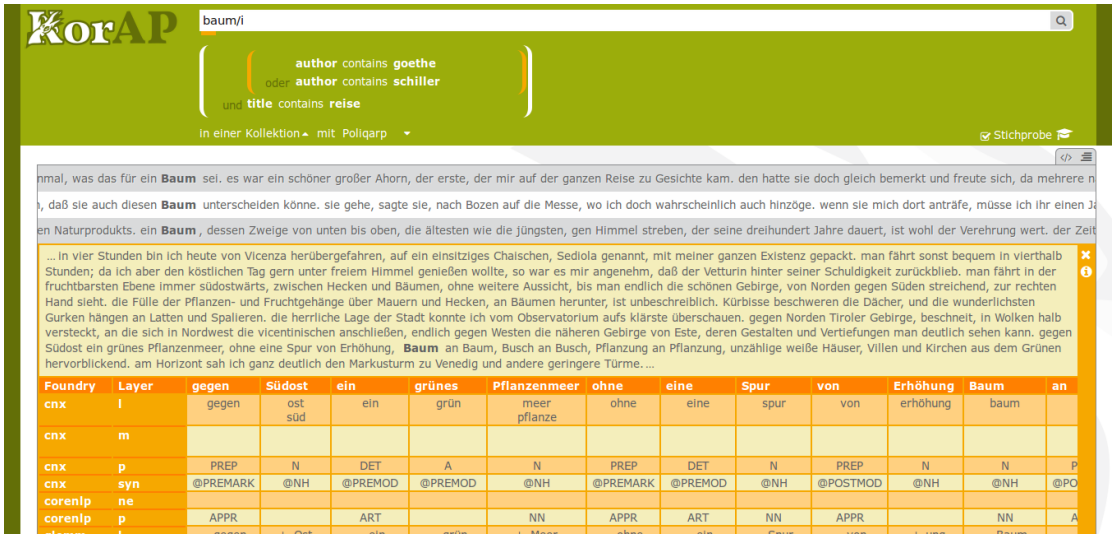

Abbildung 4: Korpussuche mit Kalamar: Anfragen (optional eingeschränkt auf eine virtuelle Kollektion) werden mit einer KWIC-Ansicht beantwortet, die Zugriff auf zugrundeliegende Annotationen erlaubt.

5. Anwendungsperspektiven

KorAP wird, nachdem es im Frühjahr 2016 in den Produktionspetrieb übergeht, über einen Zeitraum von 2-4 Jahren das Vorgängersystem COSMAS II schrittweise ersetzen. Eine wichtige Aufgabe des KorAP-Nachfolgeprojekts *Korpusrecherchesysteme* wird es dann sein, die bis dahin voraussichtlich etwa 40.000 COSMAS-II-Nutzer bei dem Umstieg auf das neue System zu unterstützen. Darüber hinaus ist aufgrund der neuen Fähigkeiten von KorAP bzgl. unterstützter Anfragesprachen, der Analyse von Konstituenz- und Dependenz-Annotationen und seiner Erweiterbarkeit eine Verbreiterung des Nutzerspektrums zu erwarten.

Ab 2016 wird KorAP außerdem nicht mehr nur zur Analyse und Recherche auf den Korpora des IDS verwendet, sondern auch durch den Rat für deutsche Rechtschreibung auf seinen speziell auf Rechtschreibbeobachtung ausgerichteten standortverteilten Korpora (Fischer et al. i. V.) sowie auf dem *Reference Corpus of Contemporary Romanian Language (CoRoLa)* (siehe folgender Abschnitt). Weitere institutionelle KorAP-Nutzer sind für die kommenden Jahre zu erwarten.

6. Folgevorhaben

Als Folgevorhaben wurde das Kooperationsprojekt mit der Universität Bukarest (und assoziierten Partnern an der Rumänischen Akademie der Wissenschaften) *Sprachvergleich korpustechnologisch – Deutsch-Rumänisch* (Laufzeit 2016-2018) erfolgreich bei der Alexander-von-Humboldt-Stiftung eingeworben. Das Projekt beruht u. a. auf KorAPs Möglichkeiten der Konstruktion virtueller Korpora und der verteilten Datenhaltung sowie der entsprechenden Abbildbarkeit von Lizenzbedingungen. Es wird die Möglichkeiten kontrastiver Untersuchungen basierend auf vergleichbaren Korpora generell und im Hinblick auf spezielle linguistische Fragestellungen untersuchen.

7. Verfügbarmachung der im Projekt entwickelten Software

Um die Weiterentwicklung des KorAP-Systems – idealerweise in Kooperation mit anderen

KorAP-Nutzern – auch über die Projektzeit hinaus zu gewährleisten und eine Adaption durch vergleichbare Projekte zu erleichtern, wurde der größte Teil der KorAP-Komponenten als Open-Source unter einer freien Softwarelizenz (BSD-2-clause) veröffentlicht[12]. Weitere Komponenten sind für die Veröffentlichung vorbereitet. Die Schnittstellen, Formate und Protokolle sind an selber Stelle dokumentiert und werden stetig aktualisiert.

Um externe Beiträge zur Weiterentwicklung von KorAP besser kanalisieren und integrieren zu können und unnötige Forks zu vermeiden, wurde außerdem ein Gerrit-Code-Review eingerichtet.[13]

8. Wirtschaftliche Verwertbarkeit

Die unmittelbare wirtschaftliche Verwertbarkeit von KorAP stand nicht im Zentrum des Vorhabens. Es wurde jedoch die vereinfachte BSD-Lizenz zur Veröffentlichung der KorAP-Quellen gewählt, um etwa einer Weiterentwicklung mit kommerziellen Partnern, einer eigenen Verwertung, aber auch einer kommerziellen Verwertung durch Dritte keine unnötigen Hindernisse in den Weg zu stellen. Im Zusammenhang mit einer im weiteren Sinne eigenen Verwertung untersucht das BMBF-geförderte Verbundprojekt *Verwertung Geist* derzeit die Möglichkeiten einer Ausgründung des IDS, um korpusbasierte, sprachwissenschaftliche Ressourcen und Dienstleistungen auch für kommerzielle Interessenten anbieten zu können. Ideen zu Geschäftsmodellen, die KorAP direkt betreffen, wurden in der Pitch-Präsentation *Big Language Data for Academic and Commercial Use* (Kupietz / Belica) im Rahmen der Innovation Days 2013 vorgestellt.

9. Kooperationen

Bzgl. der Erweiterung von KorAPs Web-Service-API wurde im Rahmen des vom BMBF geförderten Verbundprojekts *Korpusbasierte Recherche und Analyse (KobRA)* insbesondere mit der Informatik-Fakultät der TU Dortmund kooperiert.

Mit dem am IDS beheimateten Projekt *Demokratiediskurs 1918-1925* wurde eine Nutzungskooperation geschlossen, um noch während der Entwicklung von KorAP Rückmeldungen hinsichtlich gewünschter Funktionen zu erhalten. Die hierfür erweiterte RAD-Umgebung (Rapid Application Development) *Rabbid* (s. Abb. 5) kann als unabhängige Komponente zur Korpus-Recherche genutzt werden (Mell / Diewald i.V.).

10. Qualifikationen des wissenschaftlichen Nachwuchses

Die Master-Arbeit von Joachim Bingel (2015), eingereicht am Institut für Computerlinguistik der Universität Heidelberg, stand in direktem Zusammenhang mit seiner Arbeit im KorAP-Projekt und wurde von Andreas Witt hauptbetreut.

11. Vorträge

13.06.2011 Bański, Piotr / Witt, Andreas: Do linguists need a corpus query lingua franca? ISO TC37 meeting. Seoul

26.10.2011 (eingeladener Vortrag). Kupietz, Marc: KorAP – Korpusanalyseplattform der nächsten Generation: Ziele und aktuelle Arbeiten, 2. Internationale Konferenz ‚Korpuslinguistik deutsch-tschechisch kontrastiv', Würzburg

[12] https://github.com/KorAP

[13] https://korap.ids-mannheim.de/gerrit/

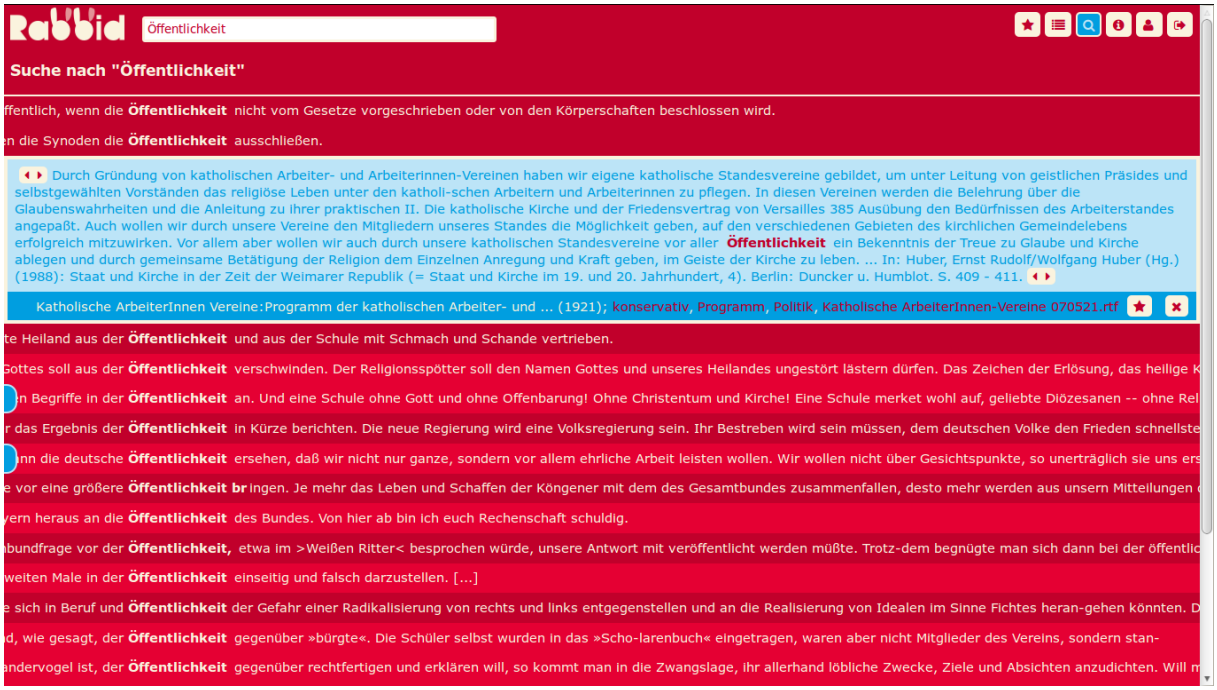

Abbildung 5: Korpussuche mit Rabbid: Rabbid ist eine *Rapid-Application-Development*-Umgebung, um neue Funktionen in KorAP (insbesondere in Kalamar) zu testen. Die Suchfunktionalität ist deutlich eingeschränkt, dafür gibt es experimentelle Funktionen wie das Kuratieren von Suchtreffern.

05.10.2012 (eingeladener Vortrag). Banski, Piotr / Frick, Elena / Schnober, Carsten / Witt, Andreas: Towards Standards for Corpus Querying, ISO Web Service Protocols Meeting, Pisa

09.11.2012 Bański, Piotr / Witt, Andreas: LingSIG: Take Three, TEI Members' Meeting, College Station, Texas, USA

10.11.2012 Bański, Piotr / Witt, Andreas: TEI for Linguists: Progress and Perspectives, TEI Members' Meeting, College Station, Texas, USA

16.11.2012 Kupietz, Marc / Witt, Andreas: KorAP – Die neue Korpusanalyseplattform, Kick-Off-Meetting des BMBF-Verbundprojekts „Korpusbasierte Recherche und Analyse (KobRA)", Dortmund

13.03.2013 Bański, Piotr / Frick, Elena / Hanl, Michael / Kupietz, Marc / Schnober, Carsten / Witt, Andreas: KorAP – Korpusanalyseplattform der nächsten Generation. 49. Jahrestagung des Instituts für Deutsche Sprache, Congress Center Rosengarten, Mannheim.

18.03.2013. Bański, Piotr / Frick, Elena / Witt, Andreas: CQLF - Corpus Query Lingua Franca: the Potsdam snapshot. ISO TC37 SC4 working group meetings, University of Potsdam.

16.7.2013. Diewald, Nils / Bański, Piotr / Witt, Andreas: Methods for Data Querying, Digital Humanities 2013, Tutorial, Lincoln, Nebraska, USA.

25.07.2013. Bański, Piotr / Frick, Elena / Hanl, Michael / Kupietz, Marc / Schnober, Carsten / Witt, Andreas: Robust corpus architecture: a new look at virtual collections and data access, Corpus Linguistics 2013, Lancaster University, http://ucrel.lancs.ac.uk/cl2013/programme.php

30.09.2013. Bański, Piotr / Kupietz, Marc / Witt, Andreas: The new corpus query engine KorAP: connections with CLARIN and the TEI, Workshop "CLARIN, Standards and the TEI", TEI Conference 2013

03.10.2013. Bański, Piotr / Witt, Andreas: LingSIG: is 4 the charm? Meeting of the TEI Special Interest Group "TEI for Linguists", TEI Conference 2013

10.12.2013. Kupietz, Marc / Belica, Cyril: Big Language Data for Academic and Commercial Use. Innovation Days, Berlin.

25.09.2014. Kupietz, Marc / Diewald, Nils: Aktuelle KobRA-relevante Entwicklungen in DeReKo und KorAP, KobRA-Workshop, Mannheim

11.3.2015. Diewald, Nils: KorAP - Architektur der Plattform. Projekttreffen ”Projekt Schreibgebrauch”, Mannheim, Germany.

11.05.2015. Kupietz, Marc (eingeladener Vortrag): Scaling out corpus technology: the open source query and analysis engine KorAP. QueryVis - Workshop on Innovative Corpus Query and Visualization Tools. NoDaLiDa. Vilnius

09.07.2015. Diewald, Nils: KorAP - Architektur der Plattform (Update). Projekttreffen ”Projekt Schreibgebrauch”, Berlin, Germany.

12. Tagungsorganisationen

22.05.2012. Workshop “Challenges in the management of large corpora CMLC-1)”, LREC 2012, Istanbul.

13.03.2013 (Marc Kupietz, Andreas Witt): ”Projektmesse: Korpora geschriebener Sprache”, IDS-Jahrestagung 2013, Mannheim

08.07.2013–10.07.2013 (Bański, Kupietz, Witt): Expertenworkshop “Perspectives for KorAP”, Sopot.

01.10.2013. Workshop “Perspectives on querying TEI-annotated data”. TEI Conference, Rom

31.05.2014. Workshop “Challenges in the management of large corpora (CMLC-2)”, LREC 2014, Reykjavik.

20.07.2015. Workshop “Challenges in the management of large corpora (CMLC-3)”, International Corpus Linguistics Conference, Lancaster.

13. Publikationen

Bański, Piotr / Biber, Hanno / Breiteneder, Evelyn / Kupietz, Marc / Lüngen, Harald / Witt, Andreas (Hgg.) (2015): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, 20 July 2015 - Mannheim: Institut für Deutsche Sprache, 2015.

Bański, Piotr / Bingel, Joachim / Diewald, Nils / Frick, Elena / Hanl, Michael / Kupietz, Marc / Pęzik, Piotr / Schnober, Carsten / Witt, Andreas (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt / Uszkoreit, Hans (Hgg.): Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference. S. 586-587 - Poznań: Fundacja Uniwersytetu im. A., 2013.

Bański, Piotr / Diewald, Nils / Hanl, Michael / Kupietz, Marc / Witt, Andreas (2014): Access Control by Query Rewriting. The Case of KorAP. In: Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14). S. 3817-3822 - European Language Resources Association (ELRA), 2014.

Bański, Piotr / Frick, Elena / Hanl, Michael / Kupietz, Marc / Schnober, Carsten / Witt, Andreas (2013): Robust corpus architecture: a new look at virtual collections and data access. In: Corpus Linguistics 2013. Abstract Book. S. 23-25 - Lancaster: UCREL, 2013.

Bański, Piotr / Kupietz, Marc / Witt, Andreas / Cavar, Damir / Heiden, Serge / Aristar, Anthony / Aristar-Dry, Helen (Hgg.) (2012): Proceedings of the LREC-2012 workshop on “Challenges in the management of large corpora” (CMLC). 22 May 2012, Istanbul, Turkey. 48 S. - European Language Resources Association (ELRA), 2012.

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: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, May 2012. S. 2905-2911 - European Language Resources Association (ELRA), 2012.

Bingel, Joachim / Diewald, Nils (2015): KoralQuery - a General Corpus Query Protocol, Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11-13, 2015, Vilnius, Litauen.

Bingel, Joachim (2015): Instantiation and Implementation of a Corpus Query Lingua Franca, Masterarbeit, Universität Heidelberg.

Diewald, Nils / Bingel, Joachim (2015): KoralQuery 0.3, Technical report, IDS, Mannheim, Working draft. http://korap.github.io/Koral/.

Fischer, Peter / Diewald, Nils / Kupietz, Marc / Witt, Andreas (i. V.): Aufbau einer Korpusinfrastruktur für die Beobachtung des Schreibgebrauchs. Abstract akzeptiert bei der Digital Humanities im deutschsprachigen Raum (DHd) 2016.

Frick, Elena / Schnober, Carsten / Bański, Piotr (2012):Evaluating Query Languages for a Corpus Processing System. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-12). Istanbul, Turkey, May 2012. S. 2286-2294 - European Language Resources Association (ELRA), 2012.

Kupietz, Marc / Lüngen, Harald / Bański, Piotr / Belica, Cyril (2014): Maximizing the Potential of Very Large Corpora. In: Kupietz, Marc / Biber, Hanno / Lüngen, Harald / Bański, Piotr / Breiteneder, Evelyn / Mörth, Karlheinz / Witt, Andreas / Takhsha, Jani (Hgg.): Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC2). S. 1-6 - European Language Resources Association (ELRA), 2014.

Kupietz, Marc / Biber, Hanno / Lüngen, Harald / Bański, Piotr / Breiteneder, Evelyn / Mörth, Karlheinz / Witt, Andreas / Takhsha, Jani (Hgg.) (2014): Proceedings of the LREC 2014 Workshop Challenges in the Management of Large Corpora (CMLC-2). 34 S. - European Language Resources Association (ELRA), 2014.

Kupietz, Marc / Frick, Elena (2013): Korpusanalyseplattform der nächsten Generation. In: Kratochvílová, Iva / Wolf, Norbert Richard (Hgg.): Grundlagen einer sprachwissenschaftlichen Quellenkunde. S. 27-36 - Tübingen: Narr, 2013. (Studien zur Deutschen Sprache 66).

Mell, Ruth / Diewald, Nils (i. V.): Korpusbasierte Diskurs-Recherche mit Rabbid. Abstract akzeptiert für das Panel „Erzählen in digitalen Diskursen: Die narrative Dimension der Neuen Medien" im Rahmen des Germanistentag 2016.

Schneider, Roman (2012): Evaluating DBMS-based Access Strategies to Very Large Multi-layer Corpora. In: Bański et al. (Hgg.): Proceedings of the LREC-12 Workshop on Challenges in the Management of Large Corpora. Istanbul, Turkey, May 2012. S. 35-48 - European Language Resources Association (ELRA), 2012.

14. Vorhabensexterne Referenzen

Bodmer, Franck (1996). Aspekte der Abfragekomponente von COSMAS II. LDV-INFO, 8:142–155. Mannheim: Institut für Deutsche Sprache.

Brouwer, Matthijs / Brugman, Hennie / Kemps-Snijders, Marc / Kunst, Jan Pieter/van der Peet, Maarten / Zeeman, Rob (i. V.), MTAS: Extending Solr and Lucene with Scalable Searchability on Annotated Text, Abstract eingereicht für die LREC 2016.

Broeder, Dan / Declerck, Thierry / Kemps-Snijders, Marc / Keibel, Holger / Kupietz, Marc / Lemnitzer, Lothar / Witt, Andreas / Wittenburg, Peter (2007): Citation of Electronic Resources: proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Dokument N366.

Evert, Stefan / Hardie, Andrew (2015): Ziggurat: A new data model and indexing format for large annotated text corpora. In Bański et al. (Hgg.): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, 20 July 2015 - Mannheim: Institut für Deutsche Sprache. S. 21 - 27.

Kupietz, Marc / Belica, Cyril / Keibel, Holger / Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta / Choukri, Khalid / Maegaard, Bente / Mariani, Joseph / Odjik, Jan / Piperidis, Stelios / Rosner, Mike / Tapias, Daniel (Hgg.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010). S. 1848-1854 - European Language Resources Association (ELRA), 2010.

Kupietz, Marc (2015): Constructing a corpus. In Durkin, Philip (Hg.): Oxford Handbook of Lexicography. Oxford: OUP.

Kupietz, Marc / Lüngen, Harald (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta et al. (Hgg.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). S. 2378-2385 - European Language Resources Association (ELRA), 2014.

Lüngen, Harald / Sperberg-McQueen, C. M. (2012): A TEI P5 Document Grammar for the IDS Text Model. In: Bański, Piotr / Modignani Picozzi, Eleonora Litta / Witt, Andreas (Hgg.): TEI and Linguistics. Journal of the Text Encoding Initiative 3. Elektronische Ressource - : 2012.

Parr, Terence J. / Quong Russell W. (1995): ANTLR: A predicated-LL (k) parser generator. Software: Practice and Experience, 25(7), S. 789-810.

Pęzik, Piotr (2013): Indexed Graph Databases for Querying Rich TEI Annotation. Conference paper presented at Perspectives on querying TEI-annotated data, TEI Conference and Members Meeting 2013, Rome, October 1, 2013.

Przepiórkowski, A. / Krynicki, Z. / Debowski, L. / Wolinski, M. / Janus, D. / Bański, P. (2004). A search tool for corpora with positional tagsets and ambiguities. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), S. 1235–1238.

Rosenfeld, V. (2010). An implementation of the Annis 2 query language. Technical report, Humboldt-Universität zu Berlin. Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. (2014). JSON-LD 1.0 – A JSON-based Serialization for Linked Data. Technical report, W3C. W3C Recommendation, http://www.w3.org/TR/json-ld/.

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

Tufiș, Dan / Mititelu, Verginica Barbu / Irimia, Elena/Ștefan Daniel Dumitrescu, Tiberiu Boroș, Horia Nicolai Teodorescu, Dan Cristea, Andrei Scutelnicu, Cecilia Bolea, Alex Moruz, Laura Pistol: CoRoLa Starts Blooming – An update on the Reference Corpus of Contemporary Romanian Language. In Bański et al. (Hgg.): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, 20 July 2015 - Mannheim: Institut für Deutsche Sprache.