GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

INSTITUT FÜR
DEUTSCHE SPRACHE

Technischer Bericht

Nr. 2016/3 (Meilenstein 4c)

Integration der KobRA-Verfahren in die IDS-Infrastrukturen

BMBF-Verbundprojekt:

Korpus-basierte linguistische Recherche und Analyse mithilfe von Data-Mining (KobRA)

Förderkennzeichen: 01UG1245D
Projektlaufzeit: 01.09.2012 bis 31.08.2015
Bearbeiter/innen: Nils Diewald, Marc Kupietz

Mannheim, den 24.2.2016

Das diesem Bericht zugrunde liegende Vorhaben wird mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter den Förderkennzeichen 01UG1245A-D gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren

Integration der KobRA-Verfahren in die IDS-Infrastrukturen

1. Bereitstellung linguistischer Ressourcen

Das Institut für Deutsche Sprache stellt mit dem Deutschen Referenzkorpus (DeReKo) das größte linguistisch motivierte elektronische Textkorpus des Gegenwartsdeutschen zur Verfügung, dessen Aufbau bereits 1964 begann (Kupietz & Keibel, 2009, Kupietz et al., 2010). Derzeit besitzt DeReKo einen Umfang von 28 Milliarden Wörtern, bei einem jährlichen Wachstum von ca. 1,7 Milliarden Wörtern (Kupietz & Lüngen, 2014). Damit verfolgt DeReKo das Hauptziel der „Maximierung von Umfang und Dispersion bzgl. potenziell relevanter Strata“ (Kupietz, 2014, S. 323).

Der Zugriff auf DeReKo erfolgt seit 2004 über die Anwendung COSMAS II (Corpus Search, Management and Analysis System, Bodmer, 1996). Seit 2011 befindet sich das Nachfolgesystem KorAP (Korpusanalyseplattform der nächsten Generation) in der Entwicklung (Bański et al., 2012, Bański et al., 2013), das mit einem Schwerpunkt auf beliebiges Datenwachstum COSMAS II in naher Zukunft ablösen wird.

Erreicht wird diese Anforderung durch horizontale Skalierbarkeit und einen hohen Grad an Modularität. Hierdurch soll zudem die Wartbarkeit und Erweiterbarkeit des Systems erleichtert werden, so dass ähnlich der Laufzeit des Vorgängersystems eine Nutzbarkeit für die nächsten 15 bis 20 Jahre erreicht werden kann.

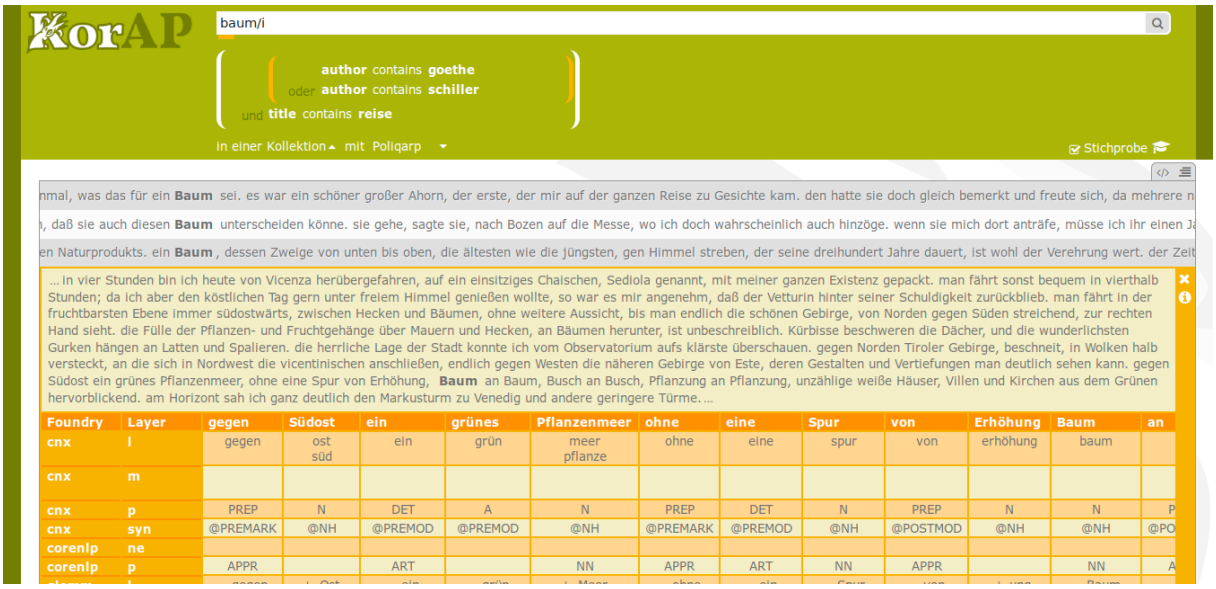

Abbildung 1: Graphische Benutzeroberfläche von KorAP.

KorAP bietet als funktionale Erweiterung von COSMAS II die Unterstützung mehrerer Anfragesprachen (Bingel & Diewald, 2015), mehrerer Annotationsschichten und eine feinere Abbildung der Nutzungsrechte von Ressourcen (Bański et al., 2014), um eine möglichst umfängliche Ausschöpfung dieser zu gewährleisten.

Seit Februar 2014 befindet sich KorAP IDS-intern im Testbetrieb. Vor dem Systemwechsel ist ein zweijähriger Parallelbetrieb zu COSMAS II geplant.

2. Integration von KobRA-Verfahren in die IDS-Infrastruktur

Aufgrund des stetigen Wachstums von DeReKo sind unterstützende Data-Mining-Verfahren besonders für die korpusbasierte Linguistik und speziell die Lexikographie von großem Interesse. Die Integration der KobRA-Verfahren in KorAP verfolgt dabei das Ziel, möglichst nutzerfreundlich und intuitiv in der Handhabung zu sein.

KorAP zielt primär auf den Einsatz als Webschnittstellen-Service ab, wobei eine grafische Nutzerschnittstelle als optionale Komponente entwickelt wurde (s. Abb. 1). Die REST-basierten API-Endpunkte und -Protokolle werden stetig erweitert und optimiert, wobei ein Hauptaugenmerk auf dem adäquaten Zugang zu Annotationen der Sprachressourcen liegt. Für die Nutzung von KobRA-Technologien bedeutet dies einen direkten Zugriff auf Ergebnislisten in Form mehrfach-annotierter XML-Fragmente oder als Bag-of-Words (s. Erfolgskontrollbericht III.2), zur Disambiguierung und zum Clustering von KWIC-Listen. Hierbei ist insbesondere die Nutzbarkeit unannotierter Rückgabedaten von Vorteil, da KorAP zwar beliebige Annotationsschichten auf Daten anbietet, jedoch nicht jedes Datum alle für eine Klassifikation benötigten Annotations-Ressourcen zur Verfügung stellen kann.

Die Kommunikationsformate sind primär JSON-basiert, wobei als zentrales Kommunikationsformat das im Projekt entwickelte JSON-LD-basierte KoralQuery eingesetzt wird (Bingel & Diewald, 2015).

KorAP unterstützt das in der CLARIN-D-Infrastruktur genutzte Single-Sing-On-Verfahren „Shibboleth“ zur Authentifizierung von Nutzern (weitere Authentifizierungssysteme werden ebenfalls unterstützt). Zur Autorisierung der Web-API (Application Programming Interface) kommt das verbreitete OAuth-2.0-Verfahren zum Einsatz, das es ermöglicht, ergänzende Funktionen, wie die Nutzung von KobRA-Verfahren, als separierte Web-Services anbieten zu können. Über die Implementation der Anfragesprache CQL und die Einrichtung der entsprechenden Schnittstelle ist KorAP in die CLARIN-Federated-Content-Search (FCS) eingebunden und stellt so die verwalteten Ressourcen einem größeren Nutzerkreis zur Verfügung. Die Anbindung an die CLARIN-Virtual-Collection-Registry ist in Vorbereitung und durch die Unterstützung dynamischer virtueller Korpora bereits grundlegend realisiert. Gleiches gilt für die Schnittstelle zu WebLicht (Hinrichs et al., 2010), die vor der Aufnahme des Parallelbetriebs noch über COSMAS II realisiert wird. Durch die Integration in die CLARIN-D-

Infrastruktur ermöglicht KorAP den erweiterten Zugriff auf die verwalteten Ressourcen und die dezentrale Nutzung weiterer KobRA-Verfahren der Projektpartner.

Zusätzlich zur dynamischen Einbindung der KobRA-Verfahren in die KorAP-Umgebung wird die automatische Textklassifikation der statischen DeReKo-Daten erprobt. Hierbei sollen sowohl neue Klassifikationen als Metadatum eingeführt als auch bestehende verbessert werden.

Um eine nachhaltige Verwendung und stetige Weiterentwicklung auch nach Projektende zu erreichen, wird KorAP als freie Software unter einer BSD-2 Lizenz veröffentlicht[1].

3. Zitierte Literatur

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C. & Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. In: Z. Vetulani & H. Uszkoreit (Hrsg.): *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.

Bański, P., Diewald, N., Hanl, M., Kupietz, M. & Witt, A. (2014). Access Control by Query Rewriting: the Case of KorAP. In: *Proceedings of the 9th conference on the Language Resources and Evaluation Conference (LREC 2014), European Language Resources Association (ELRA)* (S. 3817-3822). Reykjavic, Island, Mai 2014.

Bański, P., Fischer, P.-M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, O. & Witt, A. (2012). The New IDS Corpus Analysis Platform: Challenges and Prospects. In N. Calzolari u.a. (Hrsg.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Abgerufen unter http://www.lrec-conf.org/proceedings/lrec2012/pdf/789_Paper.pdf

Bingel, J. & Diewald, N. (2015). KoralQuery – a General Corpus Query Protocol. In: *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015* (S. 1-5), Vilnius, Litauen, 11.-13. Mai 2015.

Bodmer, F. (1996). Aspekte der Abfragekomponononente von COSMAS-II. In I. Jüttner & R. Neumann (Hrsg.), *LDV-INFO 8. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung* (S. 112–122). Mannheim: Institut für deutsche Sprache.

Diewald, N. & Bingel, J. (2015). KoralQuery 0.3. Technischer Bericht. Arbeitspapier. Abgerufen unter http://korap.github.io/Koral/

Hinrichs, E., Hinrichs, M. & Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. In J. Hajic (Hrsg.), 4*8th Annual Meeting of the Association for Computational Linguistics Proceedings of the ACL 2010 System Demonstrations.* Abgerufen unter https://aclweb.org/anthology/P/P10/P10-4005.pdf

Kupietz, M. (2014). Der Programmbereich Korpuslinguistik am IDS: Gegenwart und Zukunft. In: Ansichten und Einsichten: 50 Jahre Institut für Deutsche Sprache (S. 320-328). Institut für Deutsche Sprache, Mannheim,.

Kupietz, M. & Lüngen, H. (2014). Recent developments in DEREKO. In N. Calzolari u.a. (Hrsg.), *Proceedings of the Ninth International Conference on Language Re-sources*

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

and Evaluation (LREC'14) (S. 2378-2385). Abgerufen unter http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari u.a. (Hrsg.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Abgerufen unter http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf.

Kupietz, M. & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi & Y. Kawaguchi (Hrsg.), Working Papers in Corpus-based Linguistics and Language Education 3. Abgerufen unter http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf