Originalveröffentlichung in: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (Hrsg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. – Berlin [u.a.]: de Gruyter, 2021. S. 287-293. (Jahrbuch des Instituts für Deutsche Sprache 2020)
DOI: *https://doi.org/10.1515/9783110731514-014*

Nils Diewald (Mannheim)/Franck Bodmer (Mannheim)/
Peter Harders (Mannheim)/Elena Irimia (Bukarest)/
Marc Kupietz (Mannheim)/Eliza Margaretha (Mannheim)/
Helge Stallkamp (Mannheim)

KorAP und EuReCo – Recherchieren in mehrsprachigen vergleichbaren Korpora

Abstract: Die Korpusanalyseplattform KorAP ist von Grund auf sprachenunabhängig konzipiert. Dies gilt sowohl in Bezug auf die Lokalisierung der Benutzeroberfläche als auch hinsichtlich unterschiedlicher Anfragesprachen und der Unterstützung fremdsprachiger Korpora und ihren Annotationen. Diese Eigenschaften dienen im Rahmen der EuReCo-Initiative aktuell besonders der Bereitstellung weiterer National- und Referenzkorpora neben DeReKo. EuReCo versucht, Kompetenzen beim Aufbau großer Korpora zu bündeln und durch die Verfügbarmachung vergleichbarer Korpora quantitative Sprachvergleichsforschung zu erleichtern. Hierzu bietet KorAP inzwischen, neben dem Zugang durch die Benutzeroberfläche, einen Web API Client an, der statistische Erhebungen, auch korpusübergreifend, vereinfacht.

1 Einleitung

Seit einiger Zeit wird die Korpusrechercheplattform KorAP nicht mehr ausschließlich als Nachfolgesystem von COSMAS II (Bodmer 1996) für den Zugang zu DeReKo (Kupietz et al. 2010, 2018) eingesetzt, sondern auch für weitere National- und Referenzkorpora in Europa. Im Rahmen des Projekts (2016–2018; Cosma et al. 2017) wurde das rumänische Referenzkorpus CoRoLa (Barbu Mititelu/Tufiş/Irimia 2018) und im Rahmen des Projekts DeutUng (2017–2020) wurden Teile des ungarischen Nationalkorpus HNC (Oravecz/Váradi/Sass 2014) über KorAP zugänglich gemacht. Die Initiative, in der diese unterschiedlichen Referenzkorpora kooperieren, ist EuReCo – das „European Reference Corpus“ (Kupietz et al. 2018). Neben der Zusammenführung von Kompetenzen ist ein weiteres Ziel von EuReCo das Erstellen sogenannter „vergleichbarer Korpora“, um Sprachunterschiede und -gemeinsamkeiten in sehr großen Korpora untersuchen zu können. Hierzu werden ähnlich große Teilkorpora auf Basis von Metadaten gebildet (Bański et al. 2013), die nach

https://doi.org/10.1515/9783110731514-014

unterschiedlichen Gesichtspunkten als vergleichbar gelten können (beispielsweise hinsichtlich der Balanciertheit in Bezug auf Genres).

DRuKoLA ist ein Pilotprojekt für diese Ziele, das beweist, dass sprachunabhängige Plattformen entworfen und genutzt werden können, um europäische nationale Korpora zusammenzuführen, die vergleichende und kontrastive Sprachstudien und das Design vergleichbarer virtueller Korpora ermöglichen.

2 Sprachenunabhängigkeit

Um KorAP zu einer geeigneten Plattform für EuReCo zu machen, wurde das System von Grund auf sprachenunabhängig konzipiert. Dies betrifft sowohl die Daten- als auch die Nutzerseite.

Korpusdaten können in unterschiedlichen Sprachen vorliegen und mit beliebigen Annotationen und Metadaten angereichert werden. Da die Referenz- und Nationalkorpora, die in EuReCo gebündelt werden, in der Regel schon bestehen, wird kein spezifisches Annotations- oder Metadatenschema vorgegeben, sondern versucht, auf Basis von Grundtypen für Annotationen und Metadaten ein möglichst breites Spektrum abzudecken (Diewald/Margaretha 2017). Dies spiegelt sich auch in der Benutzeroberfläche wider, in welcher je nach hinterlegten Korpusdaten andere Musteranfragen in der Dokumentation hinterlegt werden können (siehe Abb. 1a) und unterschiedliche Annotations-Optionen im Annotations-Assistenten erscheinen (siehe Abb. 1b). Des Weiteren werden verschiedene Nutzersprachen für die Oberfläche unterstützt, um einen Einsatz in anderen Sprachräumen zu vereinfachen. Derzeit unterstützt werden Englisch, Deutsch, und in Teilen Rumänisch.

Dass die zu EuReCo gehörigen Korpora oftmals bereits existieren, bedeutet auch, dass sie bereits durch andere Korpusanalysesysteme zugänglich sind, beispielsweise durch die Corpus Workbench (CWB; Christ 1994), Annis (Zeldes et al. 2009) oder, wie im Fall von DeReKo, durch COSMAS II. Um Nutzern dieser Plattformen den Einsatz von KorAP für vergleichende Recherchen zu erleichtern, werden unterschiedliche Anfragesprachen unterstützt (derzeit die CQP-Variante Poliqarp, die COSMAS-II-Anfragesprache, ANNIS QL sowie zwei Versionen der Anfragesprache der CLARIN Federated Content Search).

Aufgrund KorAPs Eigenschaft, mit jedem beliebigen Annotations- und Metadatenschema arbeiten zu können, waren für die Integration von CoRoLa nur minimale Anpassungen erforderlich: 1. Die Konvertierung des XML-Formats der rumänischen Text- und Metadaten in das KorAP-XML-Format; 2. die Erstellung eines bestmöglichen Mappings der CoRoLa-Metadaten auf ihre DeReKo-

Entsprechung (Tufiș et al. 2019). Der zweite Schritt war insbesondere für die Erstellung virtuell vergleichbarer Korpora notwendig, da hierfür die zweistufigen Themenbereichs-Taxonomien der beiden Korpora abgebildet werden müssen (siehe Abb. 2).

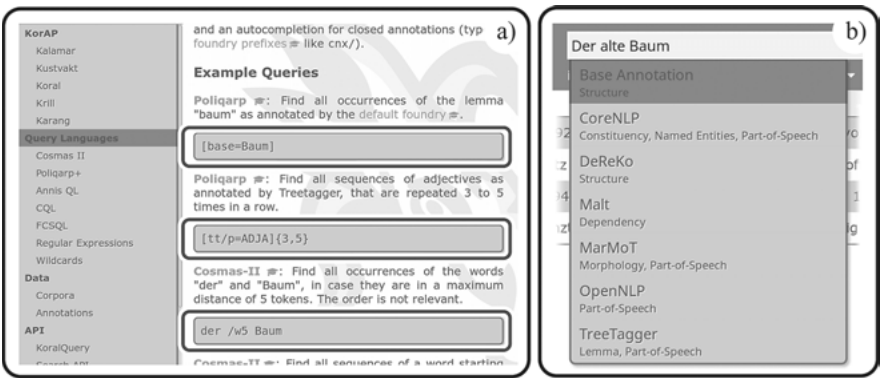

Abb. 1: Lokalisierung der Benutzeroberfläche hinsichtlich der a) Beispielanfragen in der Dokumentation und b) Annotationen im Annotations-Assistenten

Abb. 2: Parallele Anfrage an ein heterogen-annotiertes vergleichbares deutsch-rumänisches Korpus (links DeReKo--DRuKoLA-v1, rechts CoRoLa)

3 Programmierschnittstellen

Während die meisten Nutzer lediglich die grafische Webschnittstelle von KorAP kennen, existieren inzwischen einige Programme, die auf die hinterlegten Korpusdaten über eine Web-API (Application Programming Interface) zugreifen. So können Lexikon-Werkzeuge korpusbelegte Beispielsätze einbinden oder empirisch-statistische Informationen zu Wortverwendungen abfragen.

Obwohl KorAP aktuell solche statistischen Funktionen nur eingeschränkt unterstützt und über die grafische Benutzerschnittstelle zugänglich macht, ist es über diese Programmierschnittstellen schon jetzt möglich, komplexe statistische Erhebungen und Ergebnisvergleiche durchzuführen. Für die Programmiersprache R wurde hierzu eine Bibliothek namens RKorAPClient (Kupietz et al. 2020) veröffentlicht.

Abb. 3: R-Skript-Skizze zum Vergleich von Wortartenanteilen in CoRoLa, einem zu CoRoLa vergleichbaren Subkorpus von DeReKo und einem Kontroll-Subkorpus aus DeReKo

Beispielskripte, die der Bibliothek beigefügt sind, decken zahlreiche Anwendungsszenarien zur statistischen Analyse und Visualisierung von Analyseergebnissen ab und lassen sich leicht auf konkrete Anwendungen übertragen. Mit RKorAPClient lassen sich dabei nicht nur einzelne sondern auch parallel mehrere Instanzen von KorAP anfragen, um vergleichbare Korpora statistisch untersuchen zu können. So skizziert Abbildung 3, wie sich mit RKorAPClient programmatisch ein Überblick über Wortartenanteile in CoRoLa und einem mit CoRoLa vergleichbaren Subkorpus von DeReKo names DeReKo-DRuKoLA-v1 (siehe Kupietz/Cosma/Witt 2019) verschafft werden kann. Abbildung 4 vergleicht visuell die Fre-

quenzverläufe des Lemmas *Pandemie* (bzw. *pandemie*) in DeReKo-DRuKoLA-v1 und CoRoLa. Die R-Skripte zur Abfrage und zur Erzeugung der Grafiken sind ebenfalls in den Beispielskripten zum RKorAPClient enthalten.

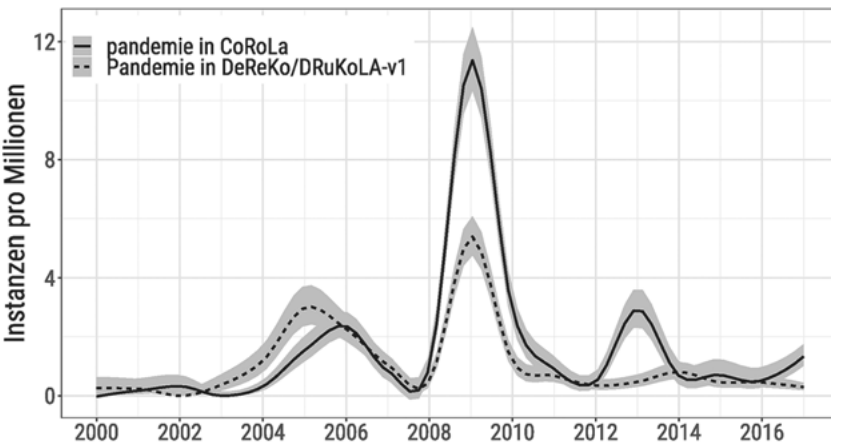

Abb. 4: Beispiel zur Nutzung der Visualisierungsfunktion des RKorAPClient mit vergleichbaren Korpora: Frequenzverläufe des Lemmas pandemie in CoRoLa und Pandemie in DeReKo-DRuKoLA-v1

4 Ausblick

Neben der nativen Unterstützung aggregierender Funktionen (bspw. Gruppierung, Sortierung), die aktuell noch durch RKorAPClient zur Verfügung gestellt werden, werden auch für unterschiedliche Anwendungszwecke zu verwendende Erweiterungen der grafischen Benutzeroberfläche für KorAP entwickelt (Diewald/Barbu Mititelu/Kupietz 2019). Auf diese Weise lassen sich Funktionen umsetzen, die nicht zur Basisfunktionalität einer Korpusanalyseplattform gehören (bspw. weil sie projekt-, ressourcen- oder sprachspezifisch sind) oder aus rechtlichen Gründen separiert entwickelt werden müssen. Dies umfasst in den nächsten Schritten unter anderem die Einbindung der aus COSMAS II bekannten Grundformensuche mit GLEMM (Belica 1994) sowie die Unterstützung verschiedener Ausgabeformate durch Export-Plugins. Aus rechtlichen Gründen und wegen der sprachlichen Spezifik (deutsche Morphologie) wird GLEMM nicht in den Kern von KorAP integriert, sondern als austauschbarer Webservice realisiert. Angeboten werden sollen wie in COSMAS II sortierte Listen von morphologisch abgeleiteten

Flexions- und Deklinationsformen, Komposita und sonstige Wortbildungsformen, die einzeln an- und abwählbar sind. Für den Export wird derzeit ein Plugin entwickelt, das als Ausgabeformat RTF und JSON anbietet. Sobald die Sortierung der Ergebnisse verfügbar ist, kann auch diese in das Export-Plugin integriert werden. Durch die Realisierung der Exportfunktionalität als Plugin ist es möglich, weitere Ausgabeformate mit Hilfe zusätzlicher, auch projektbezogener, Export-Plugins hinzuzufügen. Für EuReCo wird die Zugänglichmachung weiterer europäischer National- oder Referenzkorpora über KorAP angestrebt, um die Anzahl möglicher Paare zum Sprachvergleich stetig zu erhöhen.

Literatur

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: Hardie, Andrew/Love, Robbie (Hg.): Corpus linguistics 2013. Abstract book. Lancaster: UCREL, S. 23–25.

Barbu Mititelu, Verginica/Tufiş, Dan/Irimia, Elena (2018): The reference corpus of the contemporary romanian language (CoRoLa). In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki. Paris: European Language Resources Association (ELRA), S. 1235–1239.

Belica, Cyril (1994): WP2 – Lemmatizer. Final report MLAP93-21/WP2. Mannheim: Institut für deutsche Sprache.

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

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Papers in computational lexicography. Complex 94, S. 22–32.

Cosma, Ruxandra/Cristea, Dan/Kupietz, Marc/Tufiş, Dan/Witt, Andreas (2016): DRuKoLA – towards contrastive German-Romanian research based on comparable corpora. In: Bański, Piotr/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Witt, Andreas (Hg.): 4th Workshop on Challenges in the Management of Large Corpora (CMLC-4). Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slowenien. Paris: European Language Resources Association (ELRA), S. 28–32.

Diewald, Nils/Barbu Mititelu, Verginica/Kupietz, Marc (2019): The KorAP user interface. Accessing CoRoLa via KorAP. In: Cosma, Ruxandra/Kupietz, Marc (Hg.): On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. (= Revue Roumaine de Linguistique 64, 3). Bukarest: Editura Academiei Române, S. 265–277.

Diewald, Nils/Margaretha, Eliza (2017): Krill: KorAP search and analysis engine. In: Journal for Language Technology and Computational Linguistics (JLCL) 31, 1, S. 73–90.

Kupietz, Marc/Cosma, Ruxandra/Witt, Andreas (2019): The DRuKoLA project. In: Cosma, Ruxandra/Kupietz, Marc (Hg.): On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. (= Revue Roumaine de Linguistique 64, 3). Bukarest: Editura Academiei Române, S. 256–263.

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 2020), Marseille. Paris: ELRA, S. 7016-7021.

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/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta. Paris: European Language Resources Association (ELRA), S. 1848–1854.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German reference corpus DeReKo: new developments – new opportunities. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki. Paris: European Language Resources Association (ELRA), S. 4353–4360.

Kupietz, Marc/Cosma, Ruxandra/Cristea, Dan/Diewald, Nils/Trawiński, Beata/Tufiș, Dan/Váradi, Tamás/Wöllstein, Angelika (2018): Recent developments in the European Reference Corpus (EuReCo). In: Granger, Sylviane/Lefer, Marie-Aude/Aguiar de Souza Penha Marion, Laura (Hg.): Book of abstract. Using Corpora in Contrastive and Translation Studies Conference (5th edition). Louvain-la-Neuve: CECL Papers 1, S. 101–103.

Oravecz, Csaba/Váradi, Tamás/Sass, Bálint (2014): The hungarian gigaword corpus. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Loftsson, Hrafn/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik/Paris: European Language Resources Association (ELRA), S. 1719–1723.

Poudat, Céline/Lüngen, Harald/Herzberg, Laura (Hg.) (i. Vorb.): Wikipedia as Corpus. Erscheint in der Serie Studies in Corpus Linguistics. Amsterdam/Philadelphia: Benjamins.

Tufiș, Dan/Barbu Mititelu, Verginica/Irimia, Elena/Păiș, Vasile/Ion, Radu/Diewald, Nils/Mitrofan, Maria/Onofrei, Mihaela (2019): Little strokes fell great oaks. Creating CoRoLa. The reference corpus of contemporary Romanian. In: Cosma, Ruxandra/Kupietz, Marc (Hg.): On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. (= Revue Roumaine de Linguistique 64, 3). Bukarest: Editura Academiei Române, S. 227–240.

Zeldes, Amir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS. A search tool for multilayer annotated corpora. In: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (Hg.): Proceedings of the Corpus Linguistics 2009 Conference, Article 358. Liverpool: University of Liverpool. Internet: http://ucrel.lancs.ac.uk/publications/cl2009/ (Stand: 04.11.2020).