Meike Meliss/Vanessa González Ribao

VERGLEICHBARE KORPORA FÜR MULTILINGUALE KONTRASTIVE STUDIEN

Herausforderungen und Desiderata

Abstract This contribution aims to show the necessity of working in the development of multilingual corpora and appropriate tools for multilingual contrastive studies. We take the corpus of the lexicographical project COMBIDIGILEX as example to show, how difficult it is to build a suitable data basis to study and compare linguistic phenomena in German, Spanish and Portuguese. Despite the availability of big reference corpora for the three languages (at least for written language), it is not able to obtain a comparable data basis from, because the mentioned corpora are created according to different requirements and they are also powered by disparate information systems and analyse tools. To break the status quo, we plead for increasing research infrastructures by means of compatible language technology and sharing data.

Keywords Corpus linguistics; comparative corpora; contrastive multilingual linguistics; language technologies

1. Einleitung

Korpusbasierte Analysemethoden stellen für alle sprachlichen Beschreibungsebenen interessante empirische Daten sowohl für einzelsprachige Analysen als auch für den multilingualen Sprachvergleich bereit (Hanks 2012). Korpusevidenz durch quantitative Daten in Verbindung mit entsprechenden Forschungsfragen und Hypothesen kann den Ausgangspunkt sowohl für kontrastiv angelegte Beschreibungen von Konvergenz und Divergenz als auch für anwendungsorientierte Studien für den L2-Erwerb bilden.

In den letzten zwei Jahrzehnten ist die Zahl der verfügbaren mehrsprachigen Korpora erheblich gestiegen. Sowohl Übersetzungskorpora (= Parallelkorpora) als auch vergleichbare Korpora ermöglichen empirisch angelegte kontrastive Studien mit unterschiedlichen Ansätzen und Perspektiven (Johansson 2007, S. 5 f.; Aijmer/Altberg (Hg.) 2013, S. 1 ff.; Szudarski 2018, S. 14; Trawiński/Kupietz 2021, S. 213 ff.; Meliss i. Dr.).

Im Hinblick auf die Entwicklung mehrsprachiger vergleichbarer Korpora sind internationale Initiativen wie die Entwicklung des „International Comparable Corpus“ (ICC), das derzeit 12 Sprachen umfasst (Čermáková et al. 2021), und die Initiative zur Erstellung des „European Reference Corpus“ (EuReCo) (Kupietz et al. 2020; Diewald et al. 2021; Trawiński/Kupietz 2021) zu nennen. Der Einsatz und die Entwicklung spezifischer Analyse- und Suchinstrumente ermöglichen außerdem die Durchführung groß angelegter, mehrsprachiger kontrastiver Studien auf der Grundlage vergleichbarer empirischer Daten.

Während die deutsche Sprache in vielen der genannten Initiativen vertreten ist, gibt es bislang jedoch keine institutionellen Bestrebungen, die die Einbeziehung des Spanischen und/oder des Portugiesischen in eine der oben genannten transnationalen Projekte zur Erstellung vergleichbarer mehrsprachiger Korpora vorsehen. Ausgehend von dieser Situation ist die Durchführung von Studien mit Spanisch und Portugiesisch im Kontrast zu anderen Sprachen auf einer breiten empirischen Basis nach wie vor äußerst komplex. Die Verfügbar-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

keit einer vergleichbaren empirischen Basis ist jedoch eine der unabdingbaren Voraussetzungen für sowohl inter- als auch intralinguale Studien. Um kontrastiv angelegte empirische Studien mit dem Spanischen und/oder Portugiesischen durchzuführen, ist es daher momentan nach wie vor notwendig, *ad hoc* eine vergleichbare empirische Basis herzustellen. Dabei ist mit Johansson (2007, S. 302) zu beachten, dass die geeignete Auswahl der Sprachkorpora als empirische Grundlage unter anderem von Faktoren abhängt, die mit dem Gegenstand und dem Ziel der jeweiligen Forschungsstudie und den Forschungsfragen zusammenhängen.

Das Projekt COMBIDIGILEX[1], welches den Forschungshintergrund dieses Beitrages bildet, verfolgt u. a. das Ziel, eine geeignete Methodik für die Erstellung von korpusbasierten Studien im multilingualen Kontext (z. Z. Deutsch, Spanisch, Portugiesisch) zu entwickeln, die es ermöglicht, Forschungsfragen bezüglich konvergenter und divergenter Informationen zu dem verbalen Kombinationspotenzial im Sprachkontrast durch feingranulare Untersuchungen herauszuarbeiten. Entsprechende Pilotstudien zeigen die Möglichkeiten und Grenzen der entwickelten Methodik auf (Meliss et al. (Hg.) in Vorb.) und bilden außerdem die Datengrundlage für die Entwicklung des digitalen, multilingualen, lexiko-grammatischen Informationssystems CombiDigiLex (Fernández Méndez/Mas Álvarez/Meliss 2022). Die theoretischen und methodologischen Grundlagen des Projekts verbinden korpusbasierte Analyseansätze zum verbalen Kombinationspotenzial an der Semantik-Syntax-Schnittstelle mit semantischen Ansätzen zur Bedeutungsähnlichkeit bei Verben sowie mit der kontrastiven Linguistik im deutsch-iberoromanischen Bereich, der Korpuslinguistik und der modernen Internet-Lexikographie.

Ziel dieses Beitrags ist es, zum einen die Methoden vorzustellen, die bei der Erstellung der vergleichbaren korpusbasierten Datengrundlage für das erwähnte Projekt angewendet wurden, und zum anderen die zahlreichen Herausforderungen zu diskutieren, die hierbei bewältigt werden mussten (vgl. Abschn. 2). In dem abschließenden Abschnitt 3 werden Desiderata aufgezeigt, die für zukünftige korpusbasierte Studien im multilingualen Kontext mit den besagten Sprachen neue Wege aufweisen sollen.

2. Herausforderungen

Zunächst stellt sich die Frage, wie vergleichbar Korpora unterschiedlicher Sprachen sein können und wie ein hohes Maß an Vergleichbarkeit erzielt werden kann. Das multilinguale Arbeitskorpus des COMBIDIGILEX-Projekts setzt sich zusammen aus nach unterschiedlichen Filtern zusammengestellten Subkorpora großer einzelsprachiger (Referenz-)Korpora. Folgende vier Kriterien wurden dafür verfolgt (González Ribao/Meliss/Proost in Vorb.):

1) **Medialität**: Die Auswahl der Korpusdaten beschränkt sich auf die medial geschriebene Sprache.
2) **Verteilung und Zusammensetzung** der im Korpus vertretenen Textsorten: Das Korpus besteht aus den folgenden vier schriftsprachlichen Textsorten: Presse (P), Belletristik (BE), Wissenschaft (WI) und Gebrauchsliteratur (GL). Auf diese Weise kann der Einfluss

[1] Förderung: MINECO & FEDER (FFI2015-64476-P); vgl. https://combidigilex.wixsite.com/deutsch (letzter Zugang: 15-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

der jeweiligen Textsorte auf das Kombinationspotenzial der analysierten Verben untersucht und den diesbezüglich formulierten Forschungsfragen nachgegangen werden.

3) **Zeitraum**: Die chronologische Einschränkung und Abgrenzung auf die Zeitspanne 1990–2015 hat das Ziel, eine überschaubare Menge von aktuellen Daten[2] bereitzustellen.

4) **Sprachvarietät**: Eine geografisch-politische Eingrenzung auf die europäischen Sprachvarietäten des Spanischen und Portugiesischen und die areal definierte deutsche Sprachvarietät von Deutschland hat das Ziel, das Arbeitskorpus relativ klein zu halten.

Die damit verbundene Eingrenzung bzw. Abgrenzung der großen einzelsprachlichen (Referenz-)Korpora führt zu der Erstellung entsprechender einzelsprachlicher Subkorpora (vgl. Abb. 1). Die Datengrundlage für das Deutsche setzt sich aus Texten unterschiedlicher Korpora zusammen. Hiermit wurde hauptsächlich ein hoher Grad an inhaltlicher und typologischer Übereinstimmung der deutschen Textsammlung mit dem Textangebot der entsprechenden spanischen Referenzkorpora angestrebt. Mithilfe des Korpusrecherche-, Verwaltungs- und Analysesystems COSMAS II wurde aus dem Deutschen Referenzkorpus (= DeReKo, Release 2017)[3] für das Presse-Subkorpus ein virtuelles *Ad-hoc*-Korpus erstellt, das aus ausgewählten regionalen und überregionalen Zeitungen und verschiedenen Zeitschriften besteht. Es wurde berücksichtigt, dass die Auswahl an Zeitungen und Zeitschriften für das *Ad-hoc*-Korpus inhaltlich und thematisch dem Presseteil in CREA und CORPES ähnelt[4], um die Vergleichbarkeit der Materialien für beide Sprachen weitestgehend zu garantieren. Die anderen drei textsortenspezifischen Subkorpora wurden aus den Kernkorpora (KK) des 20. und des 21. Jahrhunderts des Digitalen Wörterbuchs der deutschen Sprache (= DWDS) zusammengesetzt (Geyken 2007). Die verwendeten spanischen Korpora und die DWDS-Kernkorpora für diese drei Textsorten weisen hinsichtlich der oben erwähnten Kriterien ein hohes Maß an Vergleichbarkeit auf (González Ribao 2021, S. 62 ff.). Für das Arbeitskorpus des Spanischen wurden die zwei Referenzkorpora der königlich spanischen Sprachakademie (= RAE) herangezogen und über die integrierte Suchmaschine abgefragt. CREA wurde für den Zeitabschnitt 1990–2000 und CORPES XXI für 2001–2015 genutzt. Zur Erstellung des Arbeitskorpus für das Portugiesische wurde das Referenzkorpus des zeitgenössischen Portugiesischen (= CRPC) verwendet, das über die Rechercheplattform CQPweb abgefragt werden kann (Mendes et al. 2012).

[2] Bei Projektbeginn lagen keine aktuelleren Daten vor.

[3] Vgl. Kupietz et al. (2018).

[4] Das DWDS bietet zwar auch ein Pressekorpus an, aber für die hier besagten Studien wurde aus folgenden Gründen mit DeReKo gearbeitet. Zum einen stellt das Letztere eine größere Vielfalt an Pressetexten (auch Zeitschriften) zur Verfügung. Zum anderen kann es über COSMAS II verwaltet werden, was ermöglicht, aus dem Angebot von DeReKo ein virtuelles *Ad-hoc*-Korpus zusammenzusetzen, das dem spanischen Angebot näherkommt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

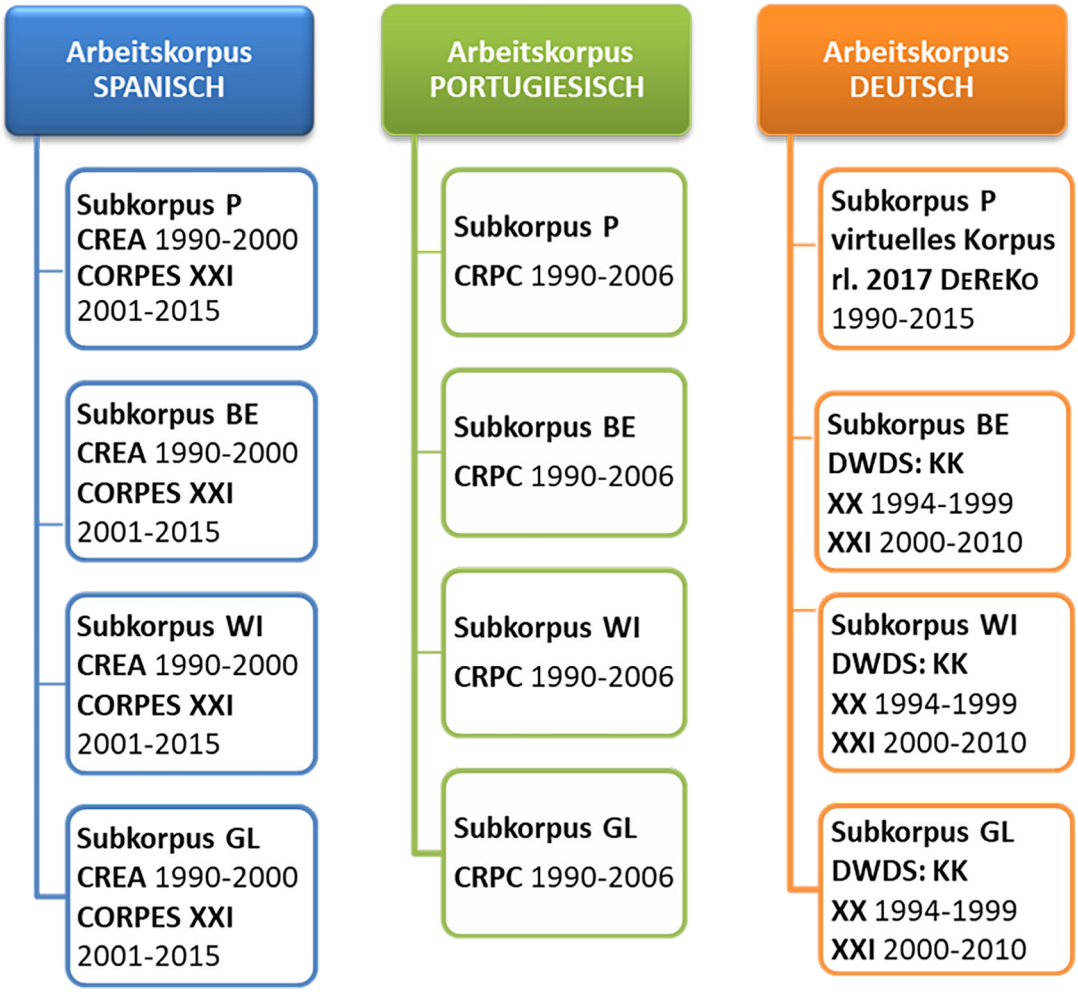

Abb. 1: Multilinguales Arbeitskorpus COMBIDIGILEX: Zusammensetzung der Subkorpora

Durch die beschriebene Methodik sollten für die drei Sprachen vergleichbare Arbeitskorpora erstellt werden, um für die projektspezifischen Forschungsfragen von COMBIDIGILEX aussagekräftige empirische Daten im multilingualen Sprachvergleich liefern zu können. Die Vergleichbarkeit der einzelnen Subkorpora konnte jedoch nur mit Einschränkung erzielt werden, weil die einzelnen Annotations-, Such- und Analysetools, über die mit den Korpora gearbeitet werden kann, nicht immer identische Funktionalitäten aufweisen. Folgende Diskrepanzen konnten aufgedeckt werden:

- **Chronologie**: Eine identische chronologische Zeitspanne konnte nicht für alle Textsorten gleichermaßen erzielt werden. Während das portugiesische Referenzkorpus auch aktuell nur Belege bis 2006 anbietet und das DWDS-Kernkorpus nur Texte bis 2010 umfasst, konnten hingegen mit dem deutschen Referenzkorpus DeReKo und dem spanischen Referenzkorpus CORPES XXI für das jeweilige Pressetextkorpus Belege bis 2015 aufgenommen werden.
- **Medialität**: Die Filterung von medial schriftlichen *vs.* medial mündlichen Texten konnte bei allen Korpora, die sowohl schriftliche als auch mündliche Daten anbieten, realisiert werden.
- **Textsorten**: Die Erstellung von textsortenspezifischen Subkorpora musste für das Deutsche durch die Kombination aus unterschiedlichen Korpora erfolgen. Da das DWDS-Zei-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tungskorpus im Gegensatz zu den spanischen und portugiesischen Referenzkorpora weniger Variation aufweist, wurde spezifisch für das deutsche Subkorpus der Pressetexte auf DeReKo zurückgegriffen. Die metadatenorientierten, textsortenspezifischen Filterfunktionen konnten in den jeweiligen Korpora allerdings zufriedenstellend angewandt werden.

- **Varietäten**: Durch eine arealorientierte Filterfunktionalität konnten in den Referenzkorpora des Spanischen und Portugiesischen die europäischen Varianten direkt gefiltert werden. Eine Beschränkung der deutschsprachigen Korpora auf den politisch-geographischen Sprachraum Deutschland konnte in DeReKo jedoch nur durch die komplexe benutzervordefinierte Auswahl der einzelnen Textkorpora erfolgen, die gleich zu Beginn der Korpusrecherche getätigt werden muss. Für die Subkorpora aus DWDS konnte keine explizite Filterfunktion bezüglich geographischer Variation genutzt werden.
- **Tools**: Die unterschiedlichen Korpusanalysetools erlauben in den meisten Fällen keinen adäquaten Export der Ergebnisse, um auf diesen eine weiterführende qualitative Analyse anzuschließen.[5] Auch die entsprechende Visualisierung der Daten im Vergleich, dem ein hoher Nutzen für die Erkenntnisgewinnung zugeschrieben werden kann, ist oft nur schwerlich zugänglich.

Die Grundlage für die **qualitativen** und **quantitativen** Analysen bildet eine entsprechende Belegsammlung, die sich auf zufallsgenerierte Stichproben von idealerweise 100 auswertbaren Belegen pro Textsorte und lexikalischer Einheit der jeweiligen Subkorpora beschränkt. Das heißt, dass nach einer ersten Bereinigung[6] angestrebt wurde, insgesamt 400 Belege pro Lexem händisch nach vorher erstellten Kodierparametern zu analysieren.[7] Die gesamte Größe der oben erwähnten Arbeitskorpora ist daher dynamisch, denn diese wachsen mit der Anzahl der Lexeme und den entsprechenden Belegsammlungen, die zur Analyse aufgenommen werden.

Für die Analysen im multilingualen Sprachvergleich sind außerdem folgende Problembereiche zu nennen:

- **Statistik**: Die Anwendung von statistischen Methoden und entsprechende Berechnungen erweisen sich bei der Arbeit mit unterschiedlich großen Korpora oft als sehr komplex (Szudarski 2018, S. 26 f.). Hinzu kommen Probleme zur Beschaffung von quantitativen Daten bei der Erstellung von Subkorpora und stratifizierten Stichproben. Bei vergleichenden Studien auf der Datengrundlage von sehr unterschiedlich großen (Teil-)Korpora ist es zudem notwendig, verschiedene Vergleichsmaße zur Berechnung anzusetzen.
- **Manuelle Analysen**: Die immer noch sehr aufwändigen manuellen einzelsprachlichen und mehrsprachigen vergleichenden Analysearbeiten erweisen sich oft als Sisyphusarbeit. Korpusbasierte und statistische Methoden erleichtern zwar unbestreitbar die Arbeit

[5] Als besonders problematisch hat sich im Fall der spanischen Korpora die Zufallsgenerierung von Samples und dessen Export sowie die quantitativen Informationen bezüglich des gesamten Korpusumfanges erwiesen. Außerdem erlaubt das entsprechende Verwaltungssystem keine Sortierung der Treffer nach dem Zufallsprinzip.

[6] Durch eine manuelle Bereinigung wurden bestimmte Belege als ungültig kodiert. Dazu wurden Merkmale wie u. a. Unvollständigkeit herangezogen. Für einige Lexeme konnten nicht immer 100 gültige Belege pro Textsorte registriert werden.

[7] Die Kodierparameter werden in González Ribao/Meliss/Proost (in Vorb.) ausführlich vorgestellt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

durch Vorstrukturierung von Massendaten und das Erkennen von bestimmten Gebrauchsphänomenen, die linguistische Interpretation bleibt jedoch nach wie vor in den Händen der LinguistInnen (Ďurčo 2010, S. 120).

3. Desiderata

Aus den aufgezeigten Problemfeldern wird deutlich, dass es unabdingbar ist, sowohl größere Mengen variationsreicher Sprachdaten für die Erstellung von multilingualen Korpora unterschiedlichster Ausprägungen bereitzustellen als auch für korpusbasierte linguistische Studien im multilingualen Kontext in Zukunft noch mehr, bessere und benutzerfreundlichere digitale Korpustechnologien zu entwickeln und einzusetzen. Dies erleichtert nicht nur die Arbeit, sondern erhöht auch die Qualität der Ergebnisse und die Anzahl der korpusbasierten Analysen an sich.

Für multilinguale Korpusstudien wären u. a. die Entwicklung benutzerfreundlicher korpusunabhängiger Such- und Analysesoftware wünschenswert, mit der (Teil-)Korpora unterschiedlicher Sprachen und verschiedener medialer Formen gleichermaßen über eine einzige Benutzeroberfläche kostenfrei abgefragt werden können. Diese multifunktionalen Werkzeuge bzw. die Integration von verschiedenen Werkzeugen müsste neben entsprechenden Filterfunktionen zu Metadaten (einzelsprachlich und im multilingualen Kontrast) und weiteren klassischen Funktionen (Konkordanzen, Kollokationen etc.) auch u. a. folgende Funktionalitäten für alle integrierten Korpora vereinen:

a) einzelsprachliche und mehrsprachige Abfragen von Kookkurrenzen (n-Gramme etc.)

b) einzelsprachlich und mehrsprachige Abfrage von annotierten Korpusdaten (POS, Formen, Semantik, Syntax etc.)

c) unterschiedliche Strukturierungsmöglichkeiten der Daten (auch Möglichkeit der Zufallsgenerierung)

d) benutzerfreundliche Exportfunktionen der Daten

e) Angebot von unterschiedlichen statistischen Methoden zur Berechnung von Häufigkeiten nach verschiedenen statistischen Parametern

f) Möglichkeiten zur Visualisierung der Daten im Vergleich

Schritte in diese aufgezeigten Richtungen werden in unterschiedlichen Projekten und an unterschiedlichen Institutionen schon seit geraumer Zeit unternommen. Ein bekanntes Beispiel für eine solche fortgeschrittene Software ist Sketch Engine mit zahlreichen Funktionalitäten für Korpora vieler Sprachen und unterschiedlicher Größen (Kilgariff et al. 2014). Es steht auch zunehmend kostenfrei verfügbare Software, wie z. B. AntConc (Anthony 2022), für spezifische Forschungsfragen im multilingualen Kontext zur Verfügung.

Bezüglich der Entwicklung von modernen multifunktionalen Recherchesystemen soll an dieser Stelle außerdem auf KorAP verwiesen werden, welches nicht nur für das Deutsche Referenzkorpus genutzt wird, sondern auch für EuReCo (Kupietz et al. 2020; Diewald et al. 2021). In diesem Rahmen werden auch weitere benutzerfreundliche Tools entwickelt (Kupietz/Diewald/Margaretha 2020).

Dennoch besteht aktuell ein klarer Bedarf an weiteren mehrsprachigen Korpora unterschiedlichster Ausprägungen, die ein hohes Maß an Vergleichbarkeit gewährleisten (Trawiński/Kupietz 2021, S. 218). Außerdem plädieren wir u. a. dafür, für multilingual-korpus-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

basierte Studien mittels einer sprachübergreifenden Korpussuch- und Analyseplattform zielgerichtet mehr Sprach- und Korpustechnologie einzusetzen. Bedingung dafür ist u. a. der freie Zugriff auf die entsprechenden Korpusdaten. Konkret für den deutsch-iberoromanischen Sprachvergleich auf der Grundlage von großen Referenzkorpora sollten die genannten Desiderata unbedingt an die oben erwähnten schon existierenden europäischen Initiativen anknüpfen, da diese für kontrastive Studien eine bessere Ausgangslage zu versprechen scheinen.[8]

Durch die Verbindung von digitaler Forschungsinfrastruktur und humanen Ressourcen auf europäischer Ebene sollten somit auch in dem Bereich der multilingualen Korpuslinguistik Synergien verstärkt gefördert und erschaffen werden. Neben einer hohen Arbeitserleichterung für korpusbasierte sprachtheoretische Fragestellungen könnte v. a. die moderne Internetlexikographie von diesen Vorschlägen sowohl bei dem lexikographischen Prozess als auch bei der Einbindung der Daten in entsprechende Ressourcen für die unterschiedlichsten Zielgruppen und Benutzersituationen profitieren (Gouws 2021, S. 16).

Literatur

Aijmer, K./Altenberg, B. (Hg.) (2013): Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson. Amsterdam/Philadelphia.

AntConc: freeware corpus analysis toolkit for concordancing and text analysis. https://www.laurenceanthony.net/software/antconc/ (Stand: 15.5.2022).

Anthony, L. (2022): What can corpus software do? In: O'Keeffe, A./McCarthy, M. (Hg.): The Routledge handbook of corpus linguistics. Abingdon/New York, Chapter 9.

Čermáková, A./Jantunen, J./Jauhiainen, T./Kirk, J./Křen, M./Kupietz, M./Uí Dhonnchadha, E. (2021): The International Comparable Corpus: challenges in building multilingual spoken and written comparable corpora. In: Research in Corpus Linguistics 9 (1). Special issue "Challenges of combining structured and unstructured data in corpus development", S. 89–103.

CORPES XXI: Corpus del Español del Siglo XXI. https://www.rae.es/banco-de-datos/corpes-xxi (Stand: 15.5.2022).

COSMAS II: Corpus Search, Management and Analysis System. https://www2.ids-mannheim.de/cosmas2/uebersicht.html und https://cosmas2.ids-mannheim.de/cosmas2-web/ (Stand: 15.5.2022).

CREA: Corpus de Referencia del Español Actual. https://corpus.rae.es/creanet.html (Stand: 15.5.2022).

CRPC: Corpus de Referência do Português Contemporâneo, Lisboa: Centro de Linguística da Universidade de Lisboa.
https://clul.ulisboa.pt/projeto/crpc-corpus-de-referencia-do-portugues-contemporaneo (Stand: 15.5.2022).

COMBIDIGILEX: Projekt: Kombinatorik in lexikalisch-semantischen Paradigmen im Kontrast. Empirische Studien und Digitalisierung für den Fremdsprachenerwerb in germanisch-iberoromanischen Kontexten. https://combidigilex.wixsite.com/deutsch (Stand: 15.5.2022).

[8] An dieser Stelle ist zu bedauern, dass sich die königlich spanische Sprachakademie bis jetzt weder an die Initiativen der EFNIL (European Federation of National Institutions for Language) noch an die EuReCo-Initiative angeschlossen hat. Kontrastive, korpusbasierte Studien mit dem Spanischen sind daher nach wie vor mit großen Herausforderungen verbunden, denen sich die Sprachforschenden mit unterschiedlichen Methoden und Strategien stellen müssen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

CombiDigiLex: Digitales, multilinguales, lexiko-grammatisches Informationssystem. Prototype V.1.0.8., Santiago de Compostela: Universidade de Santiago de Compostela. http://combidigilex.usc.gal/index.php# (Stand: 15.5.2022).

DeReKo: Deutsche Referenzkorpus. https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (Stand: 15.5.2022).

Diewald, N./Bodmer, F./Harders, P./Irimia, E./Kupietz, M./Margaretha, E./Stallkamp, H. (2021): KorAP und EuReCo – Recherchieren in mehrsprachigen vergleichbaren Korpora. In: Lobin, H./Witt, A./Wöllstein, A. (Hg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston, S. 287–294.

Ďurčo, P. (2010): Extracting data from corpora statistically – pros and cons. In: Ďurčo, P. (Hg.): Feste Wortverbindungen und Lexikographie. (= Lexicographica. Series Maior 138) Berlin/New York, S. 43–48.

DWDS: Der deutsche Wortschatz von 1600 bis heute. https://www.dwds.de/ (Stand: 18.3.2022).

EFNIL: European Federation of National Institutions for Language. http://www.efnil.org/ (Stand: 16.5.2022).

Fernández Méndez, M./Mas Álvarez, I./Meliss, M. (2022): Herausforderungen bei der Erstellung der multilingualen, korpusbasierten lexikografischen Ressource CombiDigiLex. In: TEISEL. Tecnologías para la investigación en segundas lenguas, Universitat de Barcelona, 1/2022. DOI: https://doi.org/10.1344/teisel.v1.36590.

Geyken, A. (2007): The DWDS corpus: a reference corpus for the German language of the 20th century. In: Fellbaum, C. (Hg.): Collocations and idioms: linguistic, lexicographic, and computational aspects. London, S. 23–41.

González Ribao, V. (2021): Mediale Kommunikationsverben. Das Zusammenspiel von Verb- und Musterbedeutung im Sprachvergleich Deutsch-Spanisch. (= Konvergenz und Divergenz 12). Berlin/Boston.

González Ribao, V./Meliss, M./Proost, K. (in Vorb.): Argumentstrukturen kontrastiv: Methodologische Grundlagen für korpusbasierte quantitative und qualitative Studien. In: Meliss, M./Mas Álvarez, I./Sánchez Hernández, P./González Ribao, V. (Hg.): Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen. (= Konvergenz und Divergenz). Berlin/Boston.

Gouws, R. (2021): Expanding the use of corpora in the lexicographic process of online dictionaries. In: Piosik, M./Taborek, J./Woźnicka, M. (Hg.): Korpora in der Lexikographie und Phraseologie. Stand und Perspektiven. (= Lexicographica. Series Maior 160). Berlin/Boston, S. 1–19.

Hanks, P. (2012): Corpus evidence and electronic lexicography. In: Granger, S./Paquot, M. (Hg.): Electronic lexicography. Oxford, S. 57–82.

Johansson, S. (2007): Seeing through multilingual corpora. On the use of corpora in contrastive studies. (= SCL: Studies in Corpus Linguistics 26). Amsterdam/Philadelphia.

Kilgarriff, A./Baisa, V/Bušta, J./Jakubíček, M./Kovář, V./Michelfeit, J./Rychlý, P./Suchomel, V. (2014): The Sketch Engine: ten years on. In: Lexicography: Journal of ASIALEX 1 (1), S. 7–36.

KorAP: https://korap.ids-mannheim.de/ (Stand: 15.5.2022).

Kupietz, M./Diewald, N./Margaretha, E. (2020): RKorAPClient: An R Package for accessing the German Reference Corpus DeReKo via KorAP. In: Calzolari, N./Béchet, F./Blache, P./Choukri, K./Cieri, C./Declerck, T./Goggi, S./Isahara, H./Maegaard, B./Mariani, J./Mazo, H./Moreno, A./Odijk, J./Piperidis, S. (Hg.): Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), May 11–16, 2020, Palais du Pharo, Marseille. Marseille, S. 7016–7021.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kupietz, M./Diewald, N./Trawiński, B./Cosma, R./Cristea, D./Tufiş, D./Váradi, T./Wöllstein, A. (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, S./Lefer, M. (Hg.): Translating and comparing languages: corpus-based insights. (= Corpora and Language in Use, Proceedings 6). Louvain-la-Neuve, S. 257–273.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: new developments – new opportunities. In: Calzolari, N./Choukri, K./Cieri, C./Declerck, T./Goggi, S./Hasida, K./Isahara, H./Maegaard, B./Mariani, J./Mazo, H./Moreno, A./Odijk, J./Piperidis, S./Tokunaga, T. (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, S. 4353–4360.

Meliss, M. (i. Dr.): Multilinguale Studien mit vergleichbaren Korpora: Möglichkeiten, Grenzen und Desiderata für den deutsch-iberoromanischen Kontext. In: Kongressakten IVG Palermo 2021. Jahrbuch für internationale Germanistik. (= Publikationen der Internationalen Vereinigung für Germanistik). Frankfurt a. M. u. a.

Meliss, M./Mas Álvarez, I./Sánchez Hernández, P./González Ribao, V. (Hg.) (in Vorbr.): Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen. (= Konvergenz und Divergenz). Berlin/Boston.

Mendes, A./Généreux, M./Hendrickx, I./Pereira, L./Bacelar do Nascimento, M. F./Antunes, S. (2012): CQPWeb: Uma nova plataforma de pesquisa para o CRPC. In: Costa, A./Flores, C./Alexandre, N. (Hg.): XXVII Encontro Nacional da Associação Portuguesa de Linguística. Textos Seleccionados 2011. Lissabon, S. 466–477.

RAE = REAL ACADEMIA ESPAÑOLA http://www.rae.es (letzter Zugang: 15-05-2022).

Szudarski, P. (2018): Corpus linguistics for vocabulary. A guide for research. London/New York.

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

Kontaktinformationen

Meike Meliss
Universidad de Santiago de Compostela
meike.meliss@usc.es

Vanessa González Ribao
Postdoc-Stipendiatin der Fritz-Thyssen-Stiftung
vanessina_gr@hotmail.com

Danksagung

Wir danken für den finanziellen Teilsupport über die Forschungsgruppe Humboldt GI-1920 (USC). An dieser Stelle bedanken wir uns auch herzlich bei den zwei anonymen Gutachtenden für ihre hilfreichen Kommentare sowie bei dem COMBIDIGILEX-Team für die Anwendung der entwickelten korpusbasierten Methodik bei den multilingualen Analysen und die entsprechenden kritischen Rückmeldungen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.