
Multilinguale Studien mit vergleichbaren Korpora: Möglichkeiten, Grenzen und Desiderata für den deutsch- iberoromanischen Kontext

Meike Meliss (Santiago de Compostela)

Abstract: Kontrastiv-multilingual angelegte empirische Studien erfordern eine vergleichbare Datengrundlage. Je nachdem, welche Forschungsfragen im Zentrum der sprachvergleichen- den Untersuchungen stehen, bieten sich entweder Parallelkorpora oder vergleichbare einzel- sprachliche Korpora als Datengrundlage an. Dieser Beitrag verfolgt hauptsächlich das Ziel, die Herausforderungen aufzuzeigen, die die Arbeit mit vergleichbaren Korpora im multilingualen Sprachvergleich aufwirft. Dabei soll u.a. das Prinzip der Vergleichbarkeit von Korpora themati- siert und methodologische Vorschläge für konkrete empirisch angelegte sprachvergleichende Analysen vorgelegt werden. Die Möglichkeiten und Grenzen der empirisch basierten quanti- tativen und qualitativen Analysearbeit werden durch die Präsentation einiger exemplarischer Forschungsfragen und -ergebnisse aufgezeigt. Einige Desiderata für zukünftige korpusbasierte Studien auf der Basis von vergleichbaren Korpora im multilingualen Raum schließen den Bei- trag ab.

Keywords: Korpuslinguistik, vergleichende Korpora, kontrastiv-multilinguale Linguistik, Sprachtechnologien

1. Einleitung

Multilinguale korpusbasierte Forschungsfragen sind vielfältig und facetten- reich, die Auswahl der geeigneten Sprachkorpora für den Vergleich sind daher u.a. abhängig von unterschiedlichen Faktoren bezüglich des Untersu- chungsgegenstandes und des Untersuchungsziels (Johansson 2007: 302). Der Forschungshintergrund, auf dem dieser Beitrag aufbaut, ist das Projekt COM- BIDIGILEX¹. Hauptziele des besagten Projekts bestehen einerseits in der Ent- wicklung einer geeigneten Methodik für die Erstellung von korpusbasierten Studien im multilingualen Kontext, die es ermöglicht, Forschungsfragen

1 COMBIDIGILEX: „Die Kombinatorik in lexikalisch-semantischen Paradigmen im Kontrast. Empirische Studien und Digitalisierung für den Fremdsprachenerwerb in germanisch-iberoromanischen Kontexten“: FF12015-64476-P: MINECO & FEDER; Phase 1: 2016–2019/Phase 2: 2019–2020; Siehe auch die Projektwebseite: <https://combidigilex.wixsite.com/deutsch> [Stand: 31.10.2021].



bezüglich konvergenter und divergenter Informationen zu dem verbalen Kombinationspotenzial im Sprachkontrast durch feingranulare Studien herauszuarbeiten. Entsprechende Pilotstudien zeigen die Möglichkeiten und Grenzen der entwickelten Methodik auf (Meliss et al. i.D.)². Durch die kollaborative und dynamische Erstellung des digitalen, lexiko-grammatischen Informationssystems CombiDigiLex³ wird außerdem ein anwendungsorientiertes Ziel im Bereich des Fremdspracherwerbs verfolgt (Fernández Méndez/Mas Álvarez/Meliss 2022). Der Gegenstandsbereich des Projekts befasst sich mit Verballexemen ausgewählter semantischer Klassen (u.a. KOMMUNIKATION, EMOTION, TRANSFERENCEZ) der Sprachen Deutsch und Spanisch und teilweise auch des Portugiesischen. Die korpusbasierte Datengrundlage besteht aus nach unterschiedlichen Filtern zusammengestellten Subkorpora großer einzelsprachiger (Referenz-)Korpora in Verbindung mit den folgenden vier Textsorten: Presse (P), Belletristik (BE), Wissenschaft (WI) und Gebrauchsliteratur (GL). Die theoretischen und methodologischen Grundlagen des Projekts verbinden korpusbasierte Analyseansätze zum verbalen Kombinationspotenzial an der Semantik-Syntax-Schnittstelle (Verbvalenz, Argumentstrukturmustern und Argumentrealisierungsmustern) mit semantischen Ansätzen zur Bedeutungsähnlichkeit bei Verben (lexikalisch-semantische Felder, semantische Verbklassen, Frames etc.) sowie mit der kontrastiven Linguistik im deutsch-iberoromanischen Bereich, der Korpuslinguistik und der modernen Internet-Lexikographie. Eine ausführliche Darstellung zu den projektrelevanten Grundlagen können in González Ribao/Meliss (i. D.) und González Ribao/Meliss/Proost (i. D.) nachgelesen werden.

Der vorliegende Beitrag stellt in Abschnitt 2 die empirischen und methodologischen Grundlagen vor und diskutiert die Herausforderungen bei der Erstellung einer vergleichbaren korpusbasierten Datengrundlage. In Abschnitt 3 werden auf der Basis von ausgewählten Forschungsfragen und Teilergebnissen aus zwei Studien die Möglichkeiten und Grenzen von Studien mit vergleichbaren Teilkorpora aufgezeigt. In Abschnitt 4 sollen neben einem Fazit und einem Ausblick v.a. einige Desiderata aufgezeigt werden, die für zukünftige korpusbasierte Studien im multilingualen Kontext neue Wege aufweisen und resümierend dafür plädieren, für multilingual-korpusbasierte

- 2 Dieser Sammelband vereint Pilotstudien, die im Rahmen des Projekts COMBIDIGILEX auf den projektspezifischen methodologischen und theoretischen Grundlagen erstellt wurden.
- 3 Die Ressource CombiDigiLex ist unter folgendem Link kostenlos abrufbar: <http://combidigilex.usc.gal/index.php#> [Stand: 31.10.2021]. Es handelt sich um einen lexikographischen Prototypen, der sich zur Zeit der Erstellung dieses Beitrages in der ersten Erstellungsphase befindet und der durch kollaborative Arbeit des Projektteams dynamisch erweitert wird.

Studien mittels einer sprachübergreifenden Korpussuch- und Analyseplattform zielgerichtet mehr Sprach- und Korpustechnologie einzusetzen.

2. *Empirische Grundlage: Anforderungen an Korpora für den multilingualen Sprachvergleich*

2.1. *Allgemeine Anforderungen*

Für kontrastive empirische Studien, die auf der Datengrundlage vergleichbarer Korpora beruhen, stellt sich zunächst die Frage, was vergleichbare Korpora sind bzw. wie vergleichbar Korpora unterschiedlicher Sprachen sein können und wie ein hohes Maß an Vergleichbarkeit erzielt werden kann. Für die Studien im Rahmen des COMBIDIGILEX-Projekts wurden neben einigen allgemeinen Anforderungen an Korpora v.a. eine Reihe von spezifischen Anforderungen formuliert, die in direktem Zusammenhang mit den konkreten Projektzielen stehen.

Allgemeine Anforderungen an Korpora stehen in Verbindung mit der Notwendigkeit, dass die Korpusdaten jederzeit wiederbeschaffbar bzw. abrufbar sein sollten, damit die Ergebnisse überprüft werden können. Diesbezüglich sollten u.a. folgende Metainformationen für die Korpusbenutzer zur Verfügung stehen: Quelle und Erscheinungsjahr des Textes, Sprachvarietät (zumindest geographisch), Textsorte und Modalität (geschrieben vs. gesprochen) und Gesamtgröße des Korpus. Die Metadaten eines Korpus können schon bei der Suchanfrage als Filter dienen und erlauben die anschließende Kategorisierung der Suchergebnisse. Die Korpustexte sollten außerdem zumindest tokenisiert und lemmatisiert sein. Von der Qualität der lexikalischen und morphosyntaktischen Annotationen hängen unmittelbar die Präzision und der Erfolg bei der Suchanfrage ab. Außerdem sind adäquate Exportmöglichkeiten der Korpusdaten für entsprechende Analysearbeiten unabdingbar. Eine benutzerfreundliche Konsultations- und Recherche-Oberfläche ist darüber hinaus Voraussetzung für qualitative Analysen und aussagekräftige Ergebnisse im multilingualen Sprachvergleich.⁴

Bezüglich der spezifischen Anforderungen an Korpora, die für kontrastive Studien herangezogen werden können/sollen, lassen sich verschiedene Merkmale nennen. Für die Vergleichbarkeit von Korpora unterschiedlicher Sprachen, die als Grundlage für Analysen mit hoher linguistischer Qualität dienen sollen, sind allgemeine Kriterien wie die der Ausgewogenheit, des

4 Siehe dazu u.a. auch: González Ribao 2015 und Rojo 2021.

Korpusumfangs, der Repräsentativität und der originalsprachlichen Realisierung relevant. Weitere Kriterien zu der Auswahl von Korpora bzw. der Erstellung von vergleichbaren Subkorpora stehen jeweils in direkter Verbindung mit den Projektzielen.

2.2 *Das multilinguale Arbeitskorpus für COMBIDIGILEX*

Im Rahmen von COMBIDIGILEX wurde die Korpusauswahl neben der Berücksichtigung allgemeiner Kriterien außerdem in Verbindung mit folgenden Kriterien anvisiert:

- (i) Verteilung und Zusammensetzung der im Korpus vertretenen Textsorten
- (ii) erfasster Zeitraum (Datum der Texte)
- (iii) geografische Beschränkung (Sprachvarietät der Texte)

Dadurch konnte für das Deutsche, Spanische und Portugiesische ein multilinguales Arbeitskorpus für die geschriebene Sprache der Gegenwart in großer Reichweite in Verbindung mit verschiedenen Textsorten erstellt werden. Die damit verbundene Eingrenzung bzw. Abgrenzung der großen einzelsprachlichen (Referenz-)Korpora führt zu der Erstellung entsprechender virtueller einzelsprachlicher Subkorpora. Diese bestehen aus vier schriftlichen Textsorten: Presse, Belletristik, akademisch-wissenschaftliche Texte und Gebrauchsliteratur. Eine anvisierte chronologische Abgrenzung erfolgte durch die Zeitspanne 1990–2015.⁵ Eine geographisch-politische Eingrenzung auf die europäischen Sprachvarietäten des Spanischen und Portugiesischen und die deutsche Sprachvarietät von Deutschland hatte zunächst das Ziel, das Arbeitskorpus relativ klein zu halten (Meliss/González Ribao 2022). Im Folgenden werden die virtuellen Subkorpora der drei Sprachen kurz vorgestellt.

2.2.1 *Das virtuelle Arbeitskorpus für das Deutsche*

Das virtuelle Arbeitskorpus für das Deutsche setzt sich aus Texten unterschiedlicher Korpora zusammen. Das virtuelle Presse-Korpus wurde mit

5 Die anvisierte chronologische Eingrenzung mit Anspruch auf höchstmöglicher Aktualität konnte für die jeweiligen Sprachen nicht gleichermaßen erfolgen. Für das Spanische lagen zu dem Moment der Erstellung des multilingualen Arbeitskorpus Texte zwischen 1990 und 2000 in CREA und Texte zwischen 2001 und 2015 in CORPES vor. Für das Portugiesische konnten in dem Referenzkorpus für das zeitgenössische Portugiesisch (= CRPC) nur Texte bis 2006 abgerufen werden. Für das Deutsche standen im DWDS-Kernkorpus des 21. Jh. nur Texte bis 2010 zur Verfügung. Unter chronologischer Perspektive liegt daher hier keine absolute Vergleichbarkeit der Subkorpora vor.

dem deutschen Referenzkorpus (= DeReKo⁶) über das Korpusrecherche-, Verwaltungs- und Analyzesystem COSMAS II⁷ erstellt. COSMAS II ermöglicht neben verschiedenen Filterfunktionen die Daten nach Zufall zu sortieren und zu exportieren. Die anderen drei textsortenspezifischen Subkorpora wurden mit den Kernkorpora des 20. und des 21. Jh. des DWDS⁸ erstellt (Geyken 2007). Der Zugang zu den Daten dieser beiden Korpora erfolgt über dieselbe DWDS-Suchmaschine. Diese verfügt über zahlreiche Möglichkeiten und Filterfunktionen zur Verfeinerung der Suchanfrage. Für jede Textsorte konnten die entsprechenden Filter (Belletristik, Wissenschaft oder Gebrauchsliteratur) gesetzt werden. Chronologisch wurde die Suche im Kernkorpus des 20. Jh. auf den Zeitraum 1994 bis 1999 und im Kernkorpus des 21. Jh. notwendigerweise auf den Zeitraum 2000 bis 2010 eingeschränkt. Die Suche konnte jedoch nicht geografisch gefiltert werden. Relevant für die Zwecke des Projektes war die explizite Möglichkeit, die Daten jeweils nach Zufall zu sortieren.

2.2.2 *Das virtuelle Arbeitskorpus für das Spanische*

Für das Spanische wurden die zwei Referenzkorpora der königlich spanischen Sprachakademie (RAE) herangezogen und über die integrierte Suchmaschine abgefragt. CREA⁹ wurde für den Zeitabschnitt 1990–2000 und CORPES XXI für 2001–2015¹⁰ genutzt. Über unterschiedliche Filter konnte die Recherche chronologisch und geographisch eingeschränkt und verschiedene textsortenspezifische Subkorpora erstellt werden. Ein zufallstrefferstrukturierter Export der Belege war nicht möglich und stellte eine besondere Herausforderung bei der Erstellung des Analysesample dar.

6 Das Deutsche Referenzkorpus DEREKO bildet mit 50,6 Mrd. Wörtern (Stand 02.02.2021) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit (vgl. Webseite zu DeReKo: <https://www.ids-mannheim.de/digspra/kl/projekte/korpora/> [Stand: 31.10.2021]). Siehe dazu auch Kupietz et al. 2010 und Kupietz et al 2018.

7 COSMASII: <https://www2.ids-mannheim.de/cosmas2/uebersicht.html> [Stand: 31.10.2021] und <https://cosmas2.ids-mannheim.de/cosmas2-web/> [Stand: 31.10.2021].

8 DWDS-Kernkorpora: <https://www.dwds.de/d/korpora/kern> und <https://www.dwds.de/d/korpora/korpus21> [Stand: 31.10.2021].

9 CREA = Corpus de Referencia del Español Actual („Referenzkorpus für das Spanische der Gegenwart“): <https://corpus.rae.es/creanet.html> [Stand: 31.10.2021]. REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA). *Corpus de referencia del español actual*. <http://www.rae.es> [Stand: 31.10.2021].

10 Zu Projektbeginn waren die Korpusbelege nur bis 2015 zugänglich. Aus diesem Grunde wurde zur Erstellung der virtuellen Subkorpora diese zeitliche Begrenzung festgelegt.

2.2.3 *Das virtuelle Arbeitskorpus für das Portugiesische*

Das Referenzkorpus des zeitgenössischen Portugiesischen (= CRPC)¹¹ kann über zwei unterschiedliche Plattformen – je nach Medialität (schriftlich vs. mündlich) über die Rechercheplattform CQPweb¹² abgerufen werden (Mendes et al. 2012). Das schriftliche Subkorpus konnte über unterschiedliche Filterfunktionen auf das europäische Portugiesisch und auf die Zeitspanne 1990–2006 eingeschränkt werden und weitere projektrelevante Filter- und Exportfunktionalitäten konnten ebenfalls genutzt werden.

2.2.4 *Zwischenfazit: Wie vergleichbar sind die erstellten Subkorpora?*

Durch die beschriebene Methodik sollten für die drei Sprachen vergleichbare Subkorpora 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.

Für alle drei Sprachen stehen große Korpora zur Verfügung, die über entsprechende Such- und Analysemaschinen abgefragt werden können. Die oben aufgeführten allgemeinen Anforderungen an moderne Korpora sind daher jeweils mit Einschränkung gegeben. Als besonders problematisch haben sich jedoch z.B. im Fall der spanischen Korpora die Zufallsgenerierung von Samples und dessen Export sowie die quantitativen Informationen bezüglich des gesamten Korpusumfanges erwiesen.

Weitere Anforderungen, die projektspezifisch definiert wurden, konnten ebenfalls nur teilweise erfüllt werden, da die einzelnen Such- und Analyse-tools über die mit den Korpora gearbeitet werden kann, nicht immer identische Filtermöglichkeiten aufweisen. Es handelt sich dabei um die folgenden:

- (i) Eine identische chronologische Zeitspanne mit Anspruch auf ein Maximum an Aktualität 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

11 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: 31.10.2021]. Das Referenzkorpus des Portugiesischen der Gegenwart besteht aus 309,8 Mio. Wörtern aus Texten der Schriftsprache und aus 1,6 Mio. Wörtern aus Transkriptionen von Aufnahmen gesprochener Sprache (Stand 2012).

12 Abfragen erfolgen nach kostenloser Registrierung über die Plattform CQPweb: <http://alfc.lul.clul.pt/CQPweb/> [Stand 31.10.2021].

- DEREKO und dem spanischen Referenzkorpus CORPES XXI für das jeweilige Pressetextkorpus Belege bis 2015 aufgenommen werden.
- (ii) Die Filterung von medial schriftlichen *vs.* medial mündlichen Texten konnte für alle drei Sprachen realisiert werden. Eine deutliche Trennung des Referenzkorpus in zwei große Subkorpora schriftlich *vs.* mündlich bietet jedoch nur das Portugiesische an.
 - (iii) Die Erstellung von textsortenspezifischen Subkorpora musste für das Deutsche durch die Kombination aus unterschiedlichen Korpora erfolgen, da das Referenzkorpus DEREKO nicht die entsprechende Vielfalt anbietet. Die metadatenorientierten, textsortenspezifischen Filterfunktionen konnten in den jeweiligen Korpora allerdings zufriedenstellend angewandt werden.
 - (iv) Durch eine sprachvarietätenorientierte Filterfunktionalität konnten in den Referenzkorpora des Spanischen und Portugiesischen die europäischen Varianten direkt ausgefiltert werden. Eine geographische Beschränkung der deutschsprachigen Korpora auf den politisch-geographischen Sprachraum Deutschland konnte in DEREKO nur durch die komplexe benutzerdefinierte Auswahl der einzelnen Textkorpora erfolgen, die gleich zu Beginn der Korpusrecherche getätigt werden musste. Für die Subkorpora aus DWDS konnte keine entsprechende explizite Filterfunktion bezüglich geographischer Variation genutzt werden.
 - (v) Die unterschiedlichen Korpusanalysetools, die in den letzten Jahren Möglichkeiten zur Analyse verschiedenster grammatischer Funktionalitäten geschaffen haben, erlauben in den meisten Fällen keinen adäquaten Export der Ergebnisse, um auf diesen eine weiterführende manuelle Analyse anzuschließen.
 - (vi) Zu diesen nur angedeuteten Schwierigkeiten müssen außerdem Probleme zur Beschaffung von komplexen statistischen Informationen hinzugerechnet werden. Bei der Erstellung von textsortenspezifischen Subkorpora sollten die quantitativen Daten, über die verschiedene Frequenzwerte berechnet werden können, immer transparent bleiben, denn statistisch korrekte Informationen stehen immer in Verbindung mit der Gesamtgröße des Korpus bzw. der Teilkorpora. Bei vergleichenden Studien auf der Datengrundlage von sehr unterschiedlich großen (Teil)korpora ist es notwendig, verschiedene Vergleichsmaße zur Berechnung anzusetzen, etwa die Berechnung von Häufigkeitsklassen, wie es für den Vergleich zwischen den Korpora DEREKO und FOLK (geschrieben *vs.* gesprochen) in dem Projekt LeGeDe vorgeschlagen (Meliss et al. 2018) und durch das Tool Lexical Explorer¹³ umgesetzt wurde (Lemmenmeier-Batinić 2020).

13 Lexical Explorer: Analysetool für zwei mündliche Korpora des gesprochenen Deutschen in der Interaktion (FOLK und GeWiss) und Frequenzinformation im Vergleich

3. Korpusbasierte Datenanalyse im multilingualen Kontext

Trotz der aufgezeigten Mängel und methodologischen Unzulänglichkeiten wurde die beschriebene Datengrundlage für die empirischen Pilotstudien in COMBIDIGILEX genutzt (Meliss et al. i. D.).

Als Datengrundlage wurde auf der Basis der ausgewählten Subkorpora für jede Sprache entsprechende Samples erstellt. Diese bestehen im Idealfall aus jeweils 100 Belegen pro Lexem, Textsorte und Sprache. Für die Analyse von den Kommunikationsverben *schreiben/escribir_{sp}/escrever_{pt}* ergibt sich demnach ein Sample von 400 Belegen pro Sprache, welches sich aus den ersten 100 zufallsgenerierten Belegen jeder Textsorte zusammensetzt. Als problematisch hat sich die Tatsache erwiesen, dass nicht immer 100 Belege pro Textsorte zur Verfügung standen, wie z.B. die Studien zu den Psychverben (*sich*) *langweilen/aburrir(se)* aufzeigen konnten (Meliss/Sánchez Hernández i. D.). Die Belege wurden in ein projektspezifisch definiertes Annotationschema integriert und nach unterschiedlichen Analyseparametern händisch analysiert. Aus den gewonnenen qualitativen und quantitativen einzelsprachlichen Daten wurden entsprechende Verbprofile erstellt, die in einer abstrahierten Form als *tertium comparationis* für den Sprachvergleich dienen. Für den Sprachkontrast sind quantitative Divergenzen bezüglich der Verbprofile der Ausgangspunkt für genauere sprachvergleichende Untersuchungen. Ausgewählte Ergebnisse aus zwei Studien werden im Folgenden vorgestellt.

3.1 Studie 1

Deutsche und spanische Kommunikationsverben¹⁴ zum Ausdruck von ‚schriftlicher Kommunikation‘ am Beispiel von *schreiben/escribir* wurden ausführlich in einer Studie von Franco Barros/Meliss (i. D.) beschrieben. Die folgenden Überlegungen beziehen sich auf die Analyseergebnisse der besagten Studie. In beiden Sprachen konnten zahlreiche unterschiedliche Argumentstrukturmuster (= ASTRM) dokumentiert werden. Die qualitativen und quantitativen Daten zeigen Divergenzen in der Verteilung der Muster 1–13 in beiden Sprachen auf (vgl. Abb. 1). Kontrastive Forschungsfragen rund um die Argumentstrukturmuster sind im Rahmen von COMBIDIGILEX und spezifisch für diese ausgewählten Verben relevant und werden ansatzweise durch zwei Vergleichsstudien illustriert.

zu dem Referenzkorpus des Deutschen (DeReKo): https://www.owid.de/lexex/index_de [Stand: 31.10.2021].

14 Zu korpusbasierten Studien von Kommunikationsverben im deutsch-spanische-englischen Kontext siehe auch González Ribao (2021).

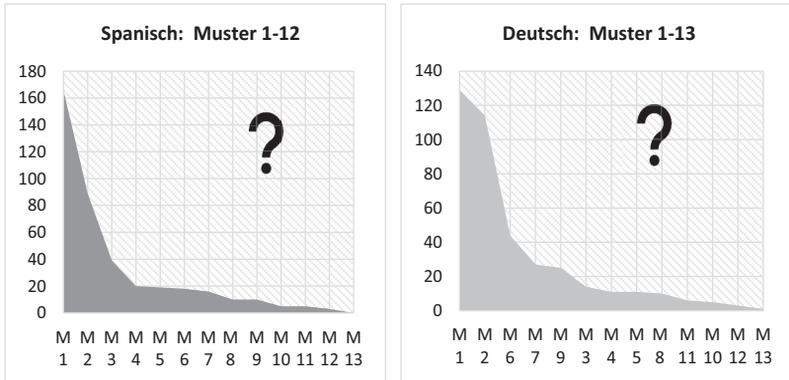


Abb. 1 a–b: Quantitative Verteilung der ASTRM für *escribir/schreiben* (absolute Zahlen).

- (i) Kontrastiv relevante Forschungsfragen befassen sich u.a. mit der Frage, wie sich die unterschiedliche quantitative Verteilung bezüglich der ASTRM erklären lässt. So ist eine quantitative Divergenz in Muster 6 besonders auffällig (vgl. Abb. 2). Warum wird M6 <jemand S *schreibt* jemandem A etwas I/T>¹⁵ im Sinne von ‚kontaktieren‘ und ‚mitteilen‘ im Spanischen seltener belegt (44 vs. 18 Belege) (vgl. Belege 1–3). Diese korpusbasierten quantitativen Evidenzen bilden den Ausgangspunkt für weitere kontrastiv angelegte Analyseschritte und Interpretationen.

- (B1) *Schiller* *fühlte sich, wie er später an Körner **schreibt**, beschämt.* (WI: DWDS 46)
- (B2) „*Ich werde ihm **schreiben**, daß wir nicht genau wußten, was wir tun sollten, und deshalb die Anzeige wiederholen, korrigiert wiederholen.*“ (BE: DWDS 110)
- (B3) *Escribidle a vuestro padre_A lo mucho que le queréis y respetáis_r* (GL-RAE 67) ‚*Schreibt eurem Vater, wie sehr ihr ihn liebt und respektiert.*‘

15 Legende: semantische Rollen: A = ADRESSAT, I = INHALT, T = THEMA/TOPIK.

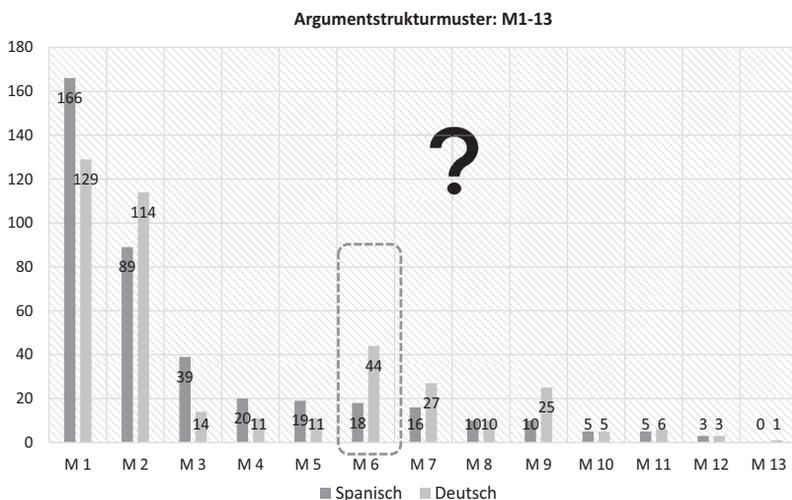


Abb. 2: Quantitative Verteilung der Argumentstrukturmuster für *escribir/schreiben* (absolute Zahlen).

Weiterhin stellt sich u.a. auch die Frage, warum z.B. die Verteilung von M6 auf die einzelnen Textsorten im Sprachvergleich so unterschiedlich ist (WI: Sp. 1 vs. Dt. 17). Gibt es vielleicht für dieses Muster im Spanischen andere Verben, die präferiert werden, wie z.B. *contar* [„erzählen“], *informar* [„informieren“/„berichten“]? (vgl. Abb. 3). Auch hier bilden die quantitativen Daten den Ausgangspunkt für weitere kontrastiv interessante Interpretationen.

Interessant ist u.a. auch die Frage, wie sich die Dativ-/Präpositivalternanz des deutschen Verbs auf spanische Entsprechungen auswirkt, da für das Spanische diese Alternanz nicht belegt werden konnte. Im Deutschen konnten für die Rolle ADRESSAT als Dativ- vs. Präpositivkomplement folgende Zahlenverhältnisse dokumentiert werden: WI: 3/17 – BE: 10/11 – GL: 7/12 – P: 1/4. Besonders auffällig ist auch hier, dass in den Texten der Wissenschaft eine deutliche Präferenz für die Realisierung durch eine Präpositionalphrase belegt wurde (WI: 3/17). Die quantitativen Daten ermöglichen den LinguistInnen einen analytischen Blick für mögliche Erklärungen.

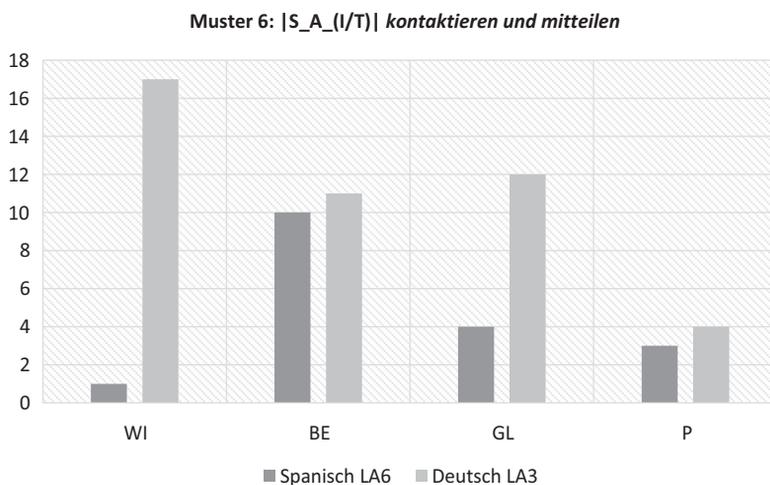


Abb. 3: Quantitative Verteilung des ASTRMs M6 auf die 4 Textsorten für *escribir/ schreiben* (absolute Zahlen). Das Muster 6 bezieht sich im Spanischen auf die Lesart 6 (LA6) und im Deutschen auf die Lesart 3 (LA3).

- (ii) Die Verteilung der Muster 1–13 in beiden Sprachen auf die vier untersuchten Textsorten (vgl. Abb. 4a–b) führt u.a. zu folgender kontrastiv orientierten Forschungsfrage: Lassen sich für die einzelnen Textsorten aus einer inter- und intralingualen Perspektive Präferenzen für das eine oder andere Muster nachweisen? Aus den Daten zu der relativen Verteilung lassen sich diesbezüglich interessante Rückschlüsse ziehen. So kann z.B. aufgezeigt werden, dass die beiden am häufigsten belegten Muster in beiden Sprachen M1 und M2 sind, dass aber in drei Textsorten (WI, BE und P) im Deutschen M2 häufiger und M1 weniger häufig als im Spanischen belegt wurde. Ausgehend von diesen quantitativen Daten kann erst eine kontrastiv angelegte Analyse der Daten eventuell genauere Interpretationen liefern.

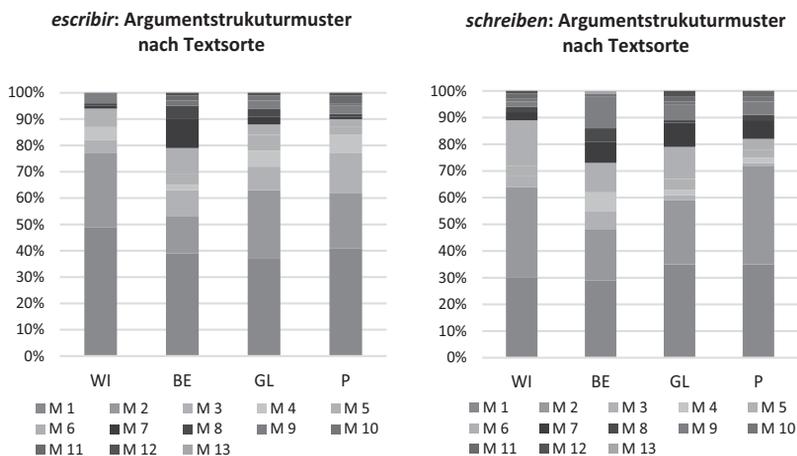


Abb. 4 a–b: Relative Verteilung der 13 ASTRM auf die 4 Textsorten (WI, BE, GL, P) für *escribir/schreiben*.

3.2 Studie 2

Ausführliche Studien zu den Psychverben (*sich*) *langweilen/aburrir(se)* im deutsch-spanischen Sprachvergleich sind von Meliss/Sánchez Hernández (i. D.) vorgelegt worden. Dieses Verbpaar weist laut der besagten Studie in beiden Sprachen drei Konstruktionsvarianten auf (vgl. Belege 4–9), die in Verbindung mit der Argumentrealisierung des Subjekts stehen. Sie belegen eine unterschiedliche Verteilungsfrequenz in beiden Sprachen.

Variante 1: EXPERIENCER als Subjekt:

- (B4) *Die Mitglieder der VS-Ortsgruppe [SUBJ_EXP] haben sich im Sommer nicht gelangweilt [..]* (P-DeReKo 45)
- (B5) *[..] Nunca me aburriré [SUBJ_EXP], porque lo bueno del cine es que cada día es diferente.* (P-RAE 286) „Ich werde mich nie langweilen, denn das Gute des Kinos ist, dass es jeden Tag anders ist.“

Variante 2: UNBELEBTEN STIMULUS als Subjekt:

- (B6) *Kunst [SUBJ_STM-1] darf alles, nur nicht langweilen [..]* (P-DeReKo 51)
- (B7) *[..] Es como si a la gente le aburriera la cultura [SUBJ_STM-1]* (P-RAE 273), Es ist so, als ob die Kultur die Leute langweilen würde.“

Variante 3: AGENTIVISCHER STIMULUS als Subjekt:

- (B8) *Und so wie ein Bettler [SUBJ_STM/AG] [..] allmählich langweilt, ärgerlich oder gar verächtlich erscheint, so wünschte auch die Mißgunst*

dem Onkel die einzige Veränderung, deren sein Leid noch fähig zu sein schien: den Tod. (BE-DWDS 73)

(B9) [. . .] *Me aburre* [SUBJ_STM/AG] *una miaja.* (BE-RAE 9) ‚Er langweilt mich ein bisschen.‘

Im Deutschen konnten außerdem auch passivische Belege dokumentiert werden, die je nach Kontext zu Variante 2 oder 3 gerechnet werden können (B10):

(B10) *Zugleich litt meine Aufmerksamkeit, aber nicht etwa, weil ich* [EXP] *gelangweilt* worden wäre, *im Gegenteil* [. . .] (BE-DWDS 69)

Eine Auswahl aus der Vielzahl von möglichen Forschungsfragen diesbezüglich sind folgende; sie werden exemplarisch in zwei Vergleichen thematisiert (vgl. Abb. 5a–b):

- Verteilen sich die drei Varianten in beiden Sprachen quantitativ ähnlich?
- Gibt es textsortenspezifische Präferenzen für die eine oder die andere Variante?
- Weisen beide Sprachen dieselbe Varianz bezüglich der Argumentstrukturmuster auf? Welche sind die häufigsten?

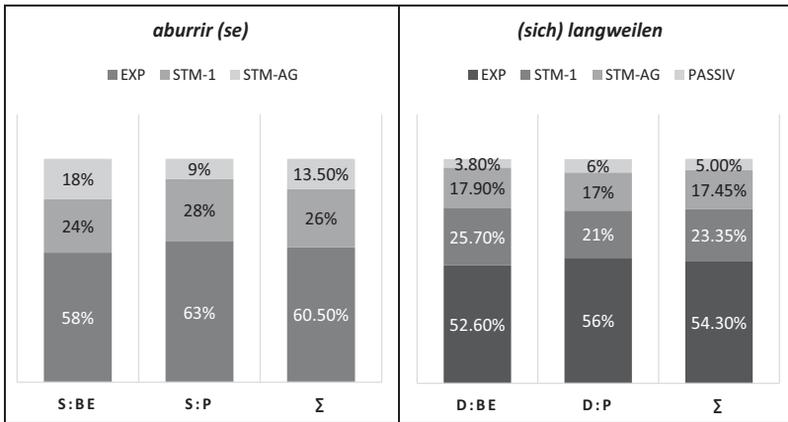


Abb. 5 a–b: Relative Verteilung der 3/4 Varianten auf die 2 Textsorten (BE und P) und die Gesamtbeleganzahl Σ für *aburrir(se)/(sich) langweilen* (relative Zahlen).

- (i) Bezüglich der Variante 1 mit einem EXPERIENCER als Subjekt lässt sich u.a. folgende kontrastive Forschungsfrage formulieren: Welche Argumentstrukturmuster lassen sich belegen? Wie ist ihre quantitative Verteilung? Aus Abb. 6 wird ersichtlich, dass in beiden Sprachen das einwertige Muster präferiert wird (vgl. Belege 11 und 12).

(B11) *Die Mitglieder der VS-Ortsgruppe haben sich im Sommer nicht gelangweilt* [. . .] (P-DeReKo 32)

(B12) [. . .] *Tampoco es que le importe: cuando Eva se aburre, se para el mundo.* (BE-RAE 705) „Das macht ihm auch nicht aus: Wenn Eva sich langweilt, stoppt die Welt.“

Die quantitativen Ergebnisse zeigen aber auch, dass weitere zwei zweiwertigen Muster in beiden Sprachen etwas häufiger als alle anderen auftreten und als solche im Spanischen zahlenmäßig etwas häufiger als im Deutschen belegt sind (Belege 13–16). An dieser quantitativen Korpusevidenz setzen dann idealerweise weitere sprachvergleichende Analyseschritte an.

Zweiwertig: SUBJ_EXPERIENCER + KMP_{SIT}–BEREICH

(B13) *Er soll sich auf keinen Fall bei uns [BEREICH] langweilen.* (BE-DWDS 104)

(B14) [. . .] *Se aburren en su isla anodina* [BEREICH]. (BE-RAE 649) „Sie langweilen sich auf ihrer belanglosen Insel.“

Zweiwertig: SUBJ_EXPERIENCER + KMP_{MOD}–ART

(B15) *Tante Else werde sein Schulgeld bezahlen, er werde wieder etwas lernen und nicht zu Hause hocken und Daumen drehen und sich zu Tode [ART] langweilen.* (BE-DWDS 59)

(B16) [. . .] *Yo me aburría como una ostra* [ART]. (P-RAE 493) „Ich langweilte mich sehr viel.“

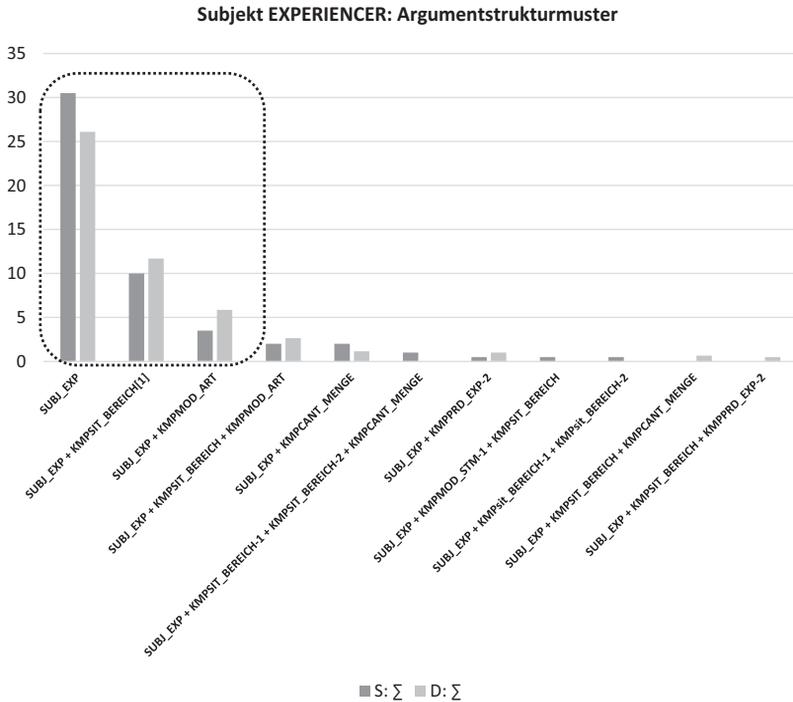


Abb. 6: ASTRM der Variante 1 und relative Verteilung bezüglich der Gesamtbeleganzahl Σ für *aburrir(se)/(sich) langweilen*.

- (ii) Für die Variante 2 mit einem unbelebten STIMULUS als Subjekt (vgl. Belege 17–20), die laut der durchgeführten Korpusanalysen im Deutschen und Spanischen ähnliche lexikalische Füller wie z.B. *Film, Wissenschaft, Politik, Erinnerungen* bzw. *discusión, datos, cultura* etc. aufweisen, lassen sich ebenfalls für beide Sprachen Forschungsfragen in Verbindung mit der quantitativen Verteilung der Argumentstrukturmuster formulieren (vgl. Abb. 7): Wie lassen sich die hohen Abweichungswerte einiger Muster im Sprachvergleich erklären? Warum weisen einige Muster nur Vorkommen in der einen oder anderen Sprache auf? Diesbezüglich ist anzumerken, dass die spanischen Psychverben auch Muster belegen können, in denen der EXPERIENCER durch ein indirektes Komplement [= In.OBJ] realisiert wird (B19–B20). Im Deutschen konnten für diese Variante des Verbs (*sich) langweilen* nur Realisierungen durch ein Akkusativkomplement [= Di.OBJ] belegt werden (B17–B18).

- (B17) *Sie hasste Politik, Politik* [STM-1] **langweilte** sie [Di.OBJ_EXP]. (BEDWDS 16)
- (B18) *Denn dieser Film* [STM-1] wird Sie [Di.OBJ_EXP] **mehr** [MENGE] **langweilen** als wiedergefundene Aldi-Kassenzettel. (P-DeReKo 66)
- (B19) [. . .] *Su actitud pusilánime* [STM-1] me [In.OBJ_EXP] **aburre** mortalmente, Elsa. (B-RAE 501) „Ihr kleinmütiges Verhalten langweilt mich zu Tode, Elsa.“
- (B20) [. . .] *Es como si a la gente le* [In.OBJ_EXP] **aburriera** la cultura [STM-1]. (P-RAE 273) „Es ist so, als ob die Kultur die Leute langweilen würde.“

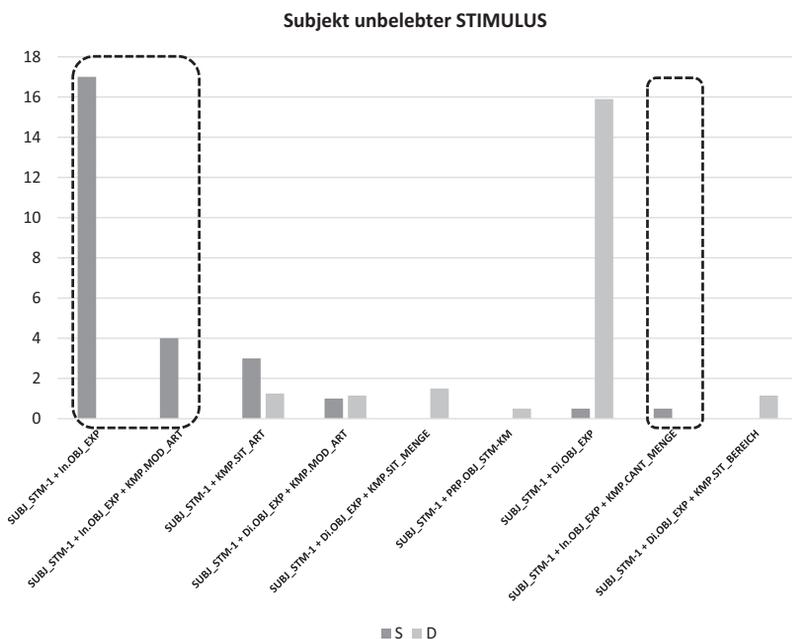


Abb. 7: ASTRM der Variante 2 und relative Verteilung bezüglich der Gesamtbeleganzahl Σ für *aburrir(se)/(sich) langweilen*.

4. Ausblick und Desiderata

Korpusbasierte Analysemethoden bieten sehr interessante empirische Daten sowohl für einzelsprachige Analysen als auch für den Sprachvergleich auf allen Beschreibungsebenen an (Hanks 2012). Korpusevidenz durch quantitative Daten kann den Ausgangspunkt für Beschreibungen von Konvergenzen und Divergenzen im Sprachvergleich bilden.

Bevor aber die Daten durch das geschulte analytische Auge von LinguistInnen genauer untersucht und beschrieben werden, lassen sich u.a. folgende Fragen bei der quantitativen Annäherung an die Daten im Vergleich anführen, deren Antworten nicht immer einfach auf der Hand liegen:

- Sind die zum Vergleich erstellten Subkorpora in ihrem Umfang wirklich miteinander vergleichbar?
- Wenn nicht, welche Vergleichsmaße können angewendet werden?
- Wie lassen sich die quantitativen Daten im Sprachvergleich statistisch korrekt berechnen und interpretieren?

Statistisch komplexe Berechnungen überfordern oft die Kompetenz von LinguistInnen, zumal die Arbeit mit unterschiedlichen Korpora, die sich alle durch unterschiedliche Größen auszeichnen, die Berechnungen entsprechend komplexer gestalten. Die Anwendung der unterschiedlichen statistischen Tests für quantitative Vergleiche wie z.B. *Log-likelihood* muss entsprechend als zusätzliche Kompetenz von LinguistInnen erlernt werden (Szudarski 2018: 26 f.).

Zusätzlich zu diesen Hürden ist zu erwähnen, dass die sehr aufwändigen manuellen einzelsprachlichen Analysearbeiten der *Sample* und den anschließenden quantitativen Evidenzen für kontrastiv relevante Fragestellungen, denen erneut qualitativ-analytisch entsprochen werden muss, zu einer Sisyphusarbeit für jeden Linguisten oder Linguistin wird. Korpusbasierte und statistische Methoden erleichtern zwar unbestreitbar die Arbeit durch Vorstrukturierung von Massendaten und Erkennen von bestimmten Gebrauchshänomenen, die linguistische Interpretation bleibt jedoch nach wie vor in den Händen der LinguistInnen (Đurčo 2010: 120). Aus diesem Grunde ist es unabdingbar, für korpusbasierte linguistische Studien im multilingualen Kontext in Zukunft noch mehr, bessere und benutzerfreundlichere digitale Technologien 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 einer benutzerfreundlichen korpusunabhängigen Such- und Analysesoftware denkbar, mit der (Teil)Korpora unterschiedlicher Sprachen und verschiedener medialer Formen (gesprochen vs. geschrieben etc.) gleichermaßen aligniert über eine einzige Benutzeroberfläche kostenfrei abgefragt werden können. Dieses multifunktionale Werkzeug bzw. die Integration von verschiedenen Werkzeugen müsste neben entsprechenden Filterfunktionen zu Metadaten (einzelsprachlich und im multilingualen Kontrast) und weiteren klassischen Basisfunktionen (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. Unterschiedliche statistische Berechnungssysteme von Häufigkeiten nach verschiedenen statistischen Parametern
- f. Möglichkeiten zur Visualisierung der Daten
- g. Kombination von Korpusrecherche- und -analysetools und Integration in lexikographische Schreibsysteme¹⁶

Neben einer hohen Arbeiterleichterung 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: 16).

Schritte in diese aufgezeigten Richtungen werden in unterschiedlichen Projekten und an unterschiedlichen Institutionen schon seit geraumer Zeit unternommen (Kilgariff/Kosen 2012). Ein Paradebeispiel für eine solche fortgeschrittene Software ist *Sketch Engine* mit zahlreichen Funktionalitäten für Lexikographen und KorpuslinguistInnen für Korpora aller Sprachen und Größen (Kilgariff et al. 2014)¹⁷. In diesem Sinne ist auch die Initiative zur Konstruktion eines internationalen vergleichbaren Korpus (= ICC) (Čermáková et al. 2021) und die zur Erstellung eines Europäischen Referenzkorpus EuReKo (Trawiński/Kupietz 2021) zu nennen. 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 EuReKo (Diewald et al. 2021; Kupietz et al. 2020)¹⁸. In diesem Rahmen werden auch weitere benutzerfreundliche Tools entwickelt (Kupietz/Diewald/Margaretha 2020).

Die dargestellten Desiderata für den deutsch-iberoromanischen Sprachvergleich auf der Grundlage von großen Referenzkorpora knüpfen an die genannten schon existierenden europäischen Initiativen an. Für die

16 Diese Forderung stellt u.a. Abel (2012: 98).

17 Da Sketch Engine nicht kostenfrei verfügbar ist, ist es nicht für alle Projektziele unbegrenzt einsetzbar.

18 An dieser Stelle soll auch auf den Beitrag von Kupietz/Trawiński (in diesen Band) verwiesen werden, in dem die vielen Vorteile von KorAP für kontrastive Korpusstudien beschrieben werden.

Verbindung von digitaler Forschungsinfrastruktur und humanen Ressourcen auf europäischer Ebene sollten auch in dem Bereich der multilingualen Korpuslinguistik Synergien gefördert und erschaffen werden.

Literaturhinweise

- Abel, Andrea (2012): Dictionary writing systems and beyond. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford University Press, 83–106.
- Čermáková, Anna/Jantunen, Jarmo/Jauhiainen, Tommi/Kirk, John/Křen, Michal/Kupietz, Marc/Uí Dhonnchadha, Elaine (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”. Murcia: Spanish Association for Corpus Linguistics, 89–103.
- Diewald, Nils/Bodmer, Franck/Harders, Peter/Irimia, Elena/Kupietz, Marc/Margaretha, Eliza/Stallkamp, Helge (2021): KorAP und EuReCo – Recherchieren in mehrsprachigen vergleichbaren Korpora. In: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (eds.): *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*. Jahrbuch des Instituts für Deutsche Sprache 2020. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston: de Gruyter, 287–294.
- Đurčo, Peter (2010): Extracting Data from Corpora Statistically – Pros and Cons. In: Đurčo, Peter (ed.): *Feste Wortverbindungen und Lexikographie*. (= *Lexicographica*. Series Maior, 138) Berlin/New York, 43–48.
- Fernández Méndez, Manuel/Mas Álvarez, Inmaculada/Meliss, Meike (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 (volumen inaugural). DOI: <https://doi.org/10.1344/teisel.v1.36590>.
- Franco Barros, Mario/Meliss, Meike (i. D.). Die Verben *schreiben/escribir* im Vergleich: empirische Studie zu Argumentstrukturmustern, Bedeutungen und Frames. In: Meliss, Meike (coord.) / Mas Álvarez, Inmaculada / Sánchez Hernández, Paloma / González Ribao, Vanessa: *Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen*. Berlin: de Gruyter [= Konvergenz und Divergenz].
- Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (ed.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London, 23–41.
- González Ribao, Vanessa/Meliss, Meike (i. D.). Argumentstrukturen in Theorie und Anwendung: Forschungsfragen im inter- und intralingualen Kontext. In: Meliss, Meike (coord.)/Mas Álvarez, Inmaculada/Sánchez Hernández, Paloma/González Ribao, Vanessa: *Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen*. Berlin: de Gruyter [= Konvergenz und Divergenz].

- González Ribao, Vanessa/Meliss, Meike/Proost, Kristel (i. D.): Argumentstrukturen kontrastiv: Methodologische Grundlagen für korpusbasierte quantitative und qualitative Studien. In: Meliss, Meike (coord.)/Mas Álvarez, Inmaculada/Sánchez Hernández, Paloma/González Ribao, Vanessa: Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen. Berlin: de Gruyter [= Konvergenz und Divergenz].
- González Ribao, Vanessa (2015): Sobre algunos conflictos en la ‚pre‘-lexicografía: la selección de corpus para la elaboración de un diccionario contrastivo alemán-español. In: Domínguez Vázquez, María José/Gómez Guinovart, Xavier/Valcárcel Riveiro, Carlos (eds.): *Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva* [Lexicography of the Romance Languages. Contemporary and Contrastive Lexicography]. Berlin/Boston: de Gruyter, 247–270.
- González Ribao, Vanessa (2021): *Mediale Kommunikationsverben. Das Zusammenspiel von Verb- und Musterbedeutung im Sprachvergleich Deutsch–Spanisch*. Berlin: de Gruyter (=Konvergenz und Divergenz, 12).
- Gouws, Rufus (2021): Expanding the use of corpora in the lexicographic process of online dictionaries. In: Piosik, Michał/Taborek, Janusz/Woźnicka, Marta (eds.): *Korpora in der Lexikographie und Phraseologie. Stand und Perspektiven*. [=Lexicographica. Series Maior, 160], Berlin/Boston: De Gruyter, 1–19.
- Hanks, Patrick (2012): Corpus evidence and electronic lexicography. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford University Press, 57–82.
- Johansson, Stig (2007): *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. (=SCL: Studies in Corpus Linguistics, 26). Amsterdam/Philadelphia: J. Benjamins.
- Kilgarriff, Adam/Kosem, Itok (2012): Corpus tools for lexicographers. In: Granger, Sylviane/Paquot, Magali (eds.): *Electronic Lexicography*. Oxford University Press, 31–55.
- Kilgarriff, Adam/Vít Baisa, Jan Bušta, Miloš Jakubiček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel (2014): The Sketch Engine: Ten Years On. In: *Lexicography: Journal of ASIALEX* Volume 1, Issue 1, 7–36.
- Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), 1848–1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf
- 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 (eds.): *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, May 11–16, 2020, Palais du Pharo, Marseille, France. Paris: European Language Resources Association, 7016–7021.

- Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, Sylviane/Lefer, Marie-Aude (eds.): *Translating and Comparing Languages: Corpus-based Insights*. (= *Corpora and Language in Use*, Proceedings 6). Louvain-la-Neuve: Presses universitaires de Louvain, 257–273.
- 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 (eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), 4353–4360.
- Lemmenmeier-Batinić, Dolores (2020): Lexical Explorer: Extending access to the Database for Spoken German for user-specific purposes. In: *Corpora* 15.1. Edinburgh: Edinburgh University Press, 55–76.
- Meliss, Meike/González Ribao, Vanessa (2022): Vergleichbare Korpora für multilinguale kontrastive Studien. Herausforderungen und Desiderata. Klosa- Kückelhaus, A. et al. (eds.): *Dictionaries and Society. Proceedings of the XX Euralex International Congress*. (12–16 July 2022, Mannheim, Germany). Mannheim, 253–261. https://euralex2022.idsmannheim.de/wp-content/uploads/2022/08/EURALEX_2022_Proceedings.pdf [Stand 24.08.2022].
- Meliss, Meike/Mas Álvarez, Inmaculada/Fernández Méndez, Manuel (coords.) (2020): CombiDigiLex. Sistema de información léxico-gramatical digital multilingüe. Prototipo V.1.0.6., Santiago de Compostela: Universidade de Santiago de Compostela: <http://combidigilex.usc.gal/index.php#> [Stand: 31.10.2021].
- Meliss, Meike/Sánchez Hernández, Paloma (i. D.). Das Konzept LANGEWEILE im Deutschen und Spanischen: Eine korpusbasierte inter- und intralinguale Studie zu verbalen Argumentstrukturmustern. In: Meliss, Meike (coord.) / Mas Álvarez, Inmaculada / Sánchez Hernández, Paloma/González Ribao, Vanessa: *Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen*. Berlin: de Gruyter [= Konvergenz und Divergenz].
- Meliss, Meike (coord.)/Mas Álvarez, Inmaculada/Sánchez Hernández, Paloma/González Ribao, Vanessa (i. D.): *Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen*. Berlin: de Gruyter [=Konvergenz und Divergenz].
- Meliss, Meike/Möhres, Christine/Batinić, Dolores/Perkuhn, Rainer (2018): Creating a list of headwords for a lexical resource of spoken German. In: Čibej, Jaka/Gorjanc, Vojko/Kosem, Iztok/Krek, Simon (eds.): *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts*, 17–21 July, Ljubljana. Slovenien: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 1009–1016.
- Mendes, Amália/Généreux, Michel/Hendrickx, Iris/Pereira, Luisa/Bacelar do Nascimento, Maria Fernanda/Antunes, Sandra (2012): CQPWeb: Uma nova plataforma de pesquisa para o CRPC. In: Costa, Armanda/Flores, Cristina/

- Alexandre, Nélia (orgs.), *XXVII Encontro Nacional da Associação Portuguesa de Linguística*. Textos Seleccionados 2011. Lisboa: APL. https://apl.pt/wp-content/uploads/2017/09/25_Mendes_et_al.pdf.
- Rojó, Guillermo (2021): *Introducción a la lingüística de corpus en español*. London: Routledge.
- Szudarski, Paweł (2018): *Corpus Linguistics for Vocabulary. A Guide for Research*. London/New York: Routledge.
- Trawiński, Beata/Kupietz, Marc (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (eds.): *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston: de Gruyter, 209–234.