Originalveröffentlichung in: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (Hrsg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. – Berlin [u.a.]: de Gruyter, 2021. S. 209-234. (Jahrbuch des Instituts für Deutsche Sprache 2020)
DOI: *https://doi.org/10.1515/9783110731514-012*

Beata Trawiński/Marc Kupietz (Mannheim)

Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo

Abstract: Der Beitrag beschreibt die Motivation und Ziele des Europäischen Referenzkorpus EuReCo, einer offenen Initiative, die darauf abzielt, dynamisch definierbare virtuelle vergleichbare Korpora auf der Grundlage bestehender nationaler, Referenz- oder anderer großer Korpora bereitzustellen und zu verwenden. Angesichts der bekannten Unzulänglichkeiten anderer Arten mehrsprachiger Korpora wie Parallel- bzw. Übersetzungskorpora oder rein webbasierte vergleichbare Korpora, stellt das EuReCo eine einzigartige linguistische Ressource dar, die neue Perspektiven für germanistische und vergleichende wie angewandte Korpuslinguistik, insbesondere im europäischen Kontext, eröffnet.

1 Einleitung

Das Europäische Referenzkorpus EuReCo (European Reference Corpus) ist eine offene Initiative, die 2013 am Leibniz-Institut für Deutsche Sprache in Mannheim ins Leben gerufen wurde. Die Hintergründe ihrer Entstehung, das zugrundeliegende Konzept und die bereits implementierten Komponenten wurden bereits auf einigen (korpus)linguistischen Konferenzen und Workshops präsentiert und in mehreren englischsprachigen Publikationen im internationalen Kontext beschrieben. Der vorliegende Beitrag fasst diese zusammen, gibt einen vollständigen Überblick über EuReCo sowie einige neuere Ergebnisse und richtet sich vor allem an die germanistische, aber auch allgemeinlinguistische Fachgemeinschaft mit sprachvergleichender und angewandter Ausrichtung.

EuReCo ist ein Vorhaben, das eine Reihe von korpustechnologischen und korpuspolitischen Herausforderungen mit sich bringt. Gleichzeitig ist es stark linguistisch motiviert und nimmt insbesondere den Sprachvergleich in den Fokus. Die beiden Aspekte werden in diesem Beitrag ausführlich diskutiert und den konzeptuellen Überlegungen hinter der EuReCo-Initiative sowie deren Umsetzung im Rahmen der zwei internationalen Projekte gegenübergestellt. Im Abschnitt 2 diskutieren wir die Anforderungen an Sprachkorpora für den Sprachvergleich und weisen auf Problematiken der vorhandenen Lösungen hin. Im Abschnitt 3 präsentieren wir die Grundidee hinter der EuReCo-Initiative und berichten über

https://doi.org/10.1515/9783110731514-012

die Ergebnisse der zwei Projekte im EuReCo-Kontext, DruKoLA und DeutUng. Abschnitt 4 beschreibt die Korpusabfrage- und Analyseplattform KorAP, die die technische Grundlage für EuReCo darstellt und den Zugang zu den im Rahmen von EuReCo erstellten vergleichbaren Korpora ermöglicht. Abschnitt 5 fasst abschließend die Resultate und Erkenntnisse der laufenden und abgeschlossenen Arbeiten im Rahmen von EuReCo zusammen, skizziert eine mögliche Weiterentwicklung von EuReCo sowie neue Perspektiven, die die EuReCo-Initiative für die germanistische und vergleichende Korpuslinguistik, insbesondere im europäischen Kontext, eröffnet.

2 Mehrsprachige Korpora für den Sprachvergleich

Es ist unumstritten, dass Korpusdaten für viele linguistische Fragestellungen substanziell sind und dass man auf Korpora in der modernen Linguistik kaum verzichten kann. Dieser Stand der Dinge hängt stark mit der empirischen Wende in der Linguistik Ende des letzten Jahrhunderts zusammen. Seitdem haben digitale Korpora sowohl in der einzelsprachlichen als auch sprachübergreifenden Forschung zunehmend an Bedeutung gewonnen. In den letzten zwei Jahrzehnten hat die Anzahl der Studien, die von Korpusdaten inspiriert sind bzw. auf Korpusdaten basieren oder auch korpusgestützt bzw. -geleitet sind, dramatisch zugenommen. Auch die Anzahl von Korpora, sowohl von monolingualen Korpora als auch bi- und multilingualen Korpora, nimmt kontinuierlich zu. Die Korpora werden auch immer größer. Als Linguist oder Linguistin steht man häufig vor dem Dilemma, aus der Vielfalt teilweise sehr unterschiedlicher Korpustypen das richtige Korpus zu wählen, wobei man bei jeder Entscheidung immer berücksichtigen muss, dass diese Auswirkungen auf die Forschungsergebnisse und letztendlich empirische und theoretische Generalisierungen haben wird. Das bezieht sich auf Fragestellungen, die sprachspezifisch sind, aber insbesondere auf Fragestellungen, die sprachübergreifende Phänomene adressieren. Denn während die Wahl eines bestimmten Korpus als Datenquelle für eine sprachspezifische Fragestellung naturgemäß auf einsprachige Korpora beschränkt ist, stehen für sprachübergreifende Fragestellungen mehrere Optionen zur Verfügung. Genauer gesagt, können sprachübergreifende Fragestellungen unter Verwendung von einsprachigen (unabhängigen) Korpora angegangen und behandelt werden oder aber auch auf der Basis von Parallelkorpora oder Vergleichskorpora untersucht werden. Im Folgenden gehen wir auf all diese Möglichkeiten ein und weisen auf ihre Vor- und Nachteile für den Sprachvergleich hin.

2.1 Einsprachige Korpora

Einsprachige Korpora sind Korpora, die Texte in nur einer Sprache enthalten und mittlerweile eine lange Tradition haben, die auf die 60er-Jahre des 20. Jahrhunderts zurückgeht. Gegenwärtig gibt es zahlreiche Korpora von vielen Einzelsprachen, darunter sehr große nationale Referenzkorpora, wie etwa das Deutsche Referenzkorpus DeReKo, die englischsprachigen Korpora American National Corpus ANC und British National Corpus BNC oder das Referenzkorpus der rumänischen Gegenwartssprache CoRoLa und das Ungarische Nationalkorpus HNC, auf die in Abschnitt 3 eingegangen wird und die aktuell eine große Rolle in EuReCo spielen.

In der Regel sind vor allem die großen nationalen Referenzkorpora lemmatisiert und morphosyntaktisch annotiert. Darüber hinaus verfügen viele monolinguale Korpora über zusätzliche linguistische Annotationen wie syntaktische Abhängigkeiten (im Sinne von Konstituentenstruktur oder Dependenzrelationen), semantische Rollen, Eigennamen, temporale Relationen, anaphorische Beziehungen, Diskurs-bezogene Relationen etc.

Monolinguale Korpora zeichnen sich schließlich durch sehr hohe und kontrollierte Sprachqualität aus, da sie (im Idealfall ausschließlich) Originaltexte enthalten und damit den muttersprachlichen Sprachgebrauch reflektieren. Gerade die hohe sprachliche Qualität ist ein entscheidender Faktor dafür, dass man auch in sprachübergreifender Forschung gerne zu monolingualen Korpora greift. Dabei werden sie meistens als Belegquellen verwendet, die einschlägige, authentische und originalsprachige Beispiele für bestimmte linguistische Phänomene oder Generalisierungen liefern. Oft werden sie aber auch für quantitative Untersuchungen verwendet. Hierzu gibt es eine Reihe von Forschungsarbeiten, einschließlich zahlreicher sprachvergleichender Studien zum Deutschen, wie z. B. Augustin (2018), Taborek (2018), Hartmann et al. (2018) und viele andere.

Während hohe sprachliche Qualität einen bedeutenden Vorteil von monolingualen Korpora darstellt, hat ihre Verwendung als Datengrundlage für sprachvergleichende Untersuchungen auch einige Defizite. Das kann anhand der Fallstudie von Hartmann et al. (2018) illustriert werden, die im Rahmen des Projekts *Deutsche Grammatik im europäischen Vergleich* (GDE) am Leibniz-Institut für Deutsche Sprache (IDS) in Mannheim durchgeführt wurde (basierend auf Wöllstein 2015 und Brandt/Trawiński/Wöllstein 2016). Der Gegenstand der Studie waren Verben im Deutschen, Schwedischen und Niederländischen, die propositionale, verbhaltige Komplemente selegieren und finite und nicht-finite Strukturen (mit und ohne Komplementierer) einbetten können. Als Beispiele im Deutschen können Verben wie *versuchen, versprechen, bitten, anordnen* etc. genannt werden. Untersucht wurde die Korrelation zwischen der Präferenz der jeweiligen

Verbtypen für finite bzw. nicht-finite Komplemente (mit und ohne Komplementierer) und den Kontrollverhältnissen bzw. (ko-)referenziellen Abhängigkeiten im Matrixsatz und eingebetteten Satz. Die Hintergrundannahme geht auf Givón (1990) und in Bezug auf das Deutsche auf Rapp et al. (2017) zurück und besagt, dass referentielle Kohäsion mit der Ereignisintegration zusammenhängt. Das Ziel war, diese Annahme für die drei germanischen Sprachen Deutsch, Schwedisch und Niederländisch anhand von Korpusdaten zu validieren. Dazu wurden drei unabhängige einsprachige Korpora herangezogen: DeReKo (Teilkorpus KoGra-DB) für das Deutsche (Kupietz et al. 2018), Språkbanken (Subcorpus Moderna) für das Schwedische (Borin et al. 2012) und LASSY Large für das Niederländische (van Noord et al. 2006, 2013). Tabelle 1 gibt einen groben Überblick über die Größe und die (Kategorien von) Texttypen in den jeweiligen Korpora.

Tab. 1: Monolinguale Korpora in Hartmann et al. (2018)

Korpus	Worttoken	Satztoken	Texttypen/Themenbereiche
DeReKo (Subkorpus KoGra-DB)	4.3 G	200 M	170 Kategorien: Presse, Roman, Gedicht, Krimi, Belletristik, Dissertation, Wettervorhersage, Werbebroschüre, Horoskop, Leserbrief, Reiseführer etc.
Språkbanken (Subkorpus Moderna)	13.3 G	953 M	Presse, Zeitschrift, Protokolle, Literatur, Bloggmix, Twittermix, Wikipedia etc.
LASSY (Korpus Large)	0.8 G	52 M	18 Kategorien: Verwaltungstexte, juristische Texte, Zeitschrift, Protokolle (Europarl), Web, Wikipedia, Thronreden der Königin Beatrix etc.

Die Distribution der relevanten Verbtypen in den drei Korpora ließ eine Korrelation zwischen Selektionspräferenzen und Kontrollverhältnissen erkennen. Damit bestätigen die Ergebnisse die Hypothese der referenziellen Kohäsion und Ereignisintegration und scheinen darüber hinaus zu zeigen, dass diese eine sprachübergreifende Gültigkeit hat.

Die methodische Frage, die sich in Zusammenhang mit dieser Studie allerdings ergibt, ist, ob bzw. inwiefern die Einzelergebnisse für das Deutsche, das Schwedische und das Niederländische miteinander vergleichbar sind. Betrachtet man die zugrundeliegende Datenbasis (Tab. 1), so muss man feststellen, dass diese sich für jede Sprache hinsichtlich der Größe und Zusammensetzung massiv unterscheidet. Ein solcher Stand der Dinge ist in Forschungsvorhaben, die auf mehreren unabhängigen einsprachigen Korpora basieren, meist der Fall.

Die Ergebnisse solcher Forschungsvorhaben sind daher auf einer Metaebene (Ebene der Generalisierungen) sicherlich vergleichbar. Auf der empirischen Ebene (Datenebene) können sie aufgrund der Datenverschiedenheit bzw. deren niedrigen Vergleichbarkeit weniger als vergleichbar gelten.

Zusammenfassend lässt sich also sagen, dass man bei einsprachigen Korpora in der Regel von niedriger Vergleichbarkeit im Sinne von Größenübereinstimmung und Übereinstimmung hinsichtlich der Komposition ausgehen muss. Gleichzeitig versprechen einsprachige Korpora relativ hohe sprachliche Qualität, die muttersprachlichen Sprachgebrauch adäquat widerspiegelt. Will man die Vergleichbarkeit und die sprachliche Qualität jeweils mit einer Skala repräsentieren, mit niedriger Vergleichbarkeit bzw. niedriger sprachlicher Qualität am linken Ende der Skala und hoher Vergleichbarkeit bzw. hoher sprachlicher Qualität am rechten Ende der Skala, dann wären monolinguale Korpora wie in Abbildung 1 charakterisierbar.

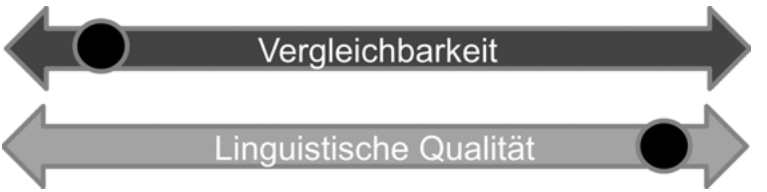

Abb. 1: Geringe Vergleichbarkeit und hohe linguistische Qualität in einsprachigen Korpora

2.2 Parallele Korpora

Hohe sprachliche Qualität ist ohne Zweifel eine bedeutende Eigenschaft einer linguistischen Datenquelle, weshalb monolinguale Korpora nicht nur in einzelsprachiger, sondern auch in sprachübergreifender Forschung gerne eingesetzt werden. Gleichzeitig stellt niedrige Vergleichbarkeit ein ernsthaftes empirisch-methodisches Problem für den Sprachvergleich dar. Aus diesem Grund greift man in sprachübergreifender Forschung doch überwiegend auf multilinguale Korpora, und insbesondere auf Parallelkorpora zurück.

Parallelkorpora bestehen aus Originaltexten in einer Sprache (Quellsprache) und ihren Übersetzungen in anderen Sprachen (Zielsprachen); daher werden sie manchmal als Übersetzungskorpora bezeichnet (das ist zum Beispiel in der Translationswissenschaft der Fall). Die Paralleltexte sind in der Regel in allen Sprachen auf Satzebene aligniert und sind linguistisch annotiert, wobei die Detailliertheit

der linguistischen Annotation von Parallelkorpus zu Parallelkorpus stark variieren kann. Parallelkorpora werden erst seit den 1990er Jahren entwickelt und setzten teilweise andere Technologien als monolinguale Korpora voraus.[1] Mittlerweile gibt es eine Reihe von elektronischen Parallelkorpora, die frei zugänglich sind und über verschiedene webbasierte Recherche- und Analysesysteme durchsucht werden können. Zu den größten gehören aktuell The Open Parallel Corpus OPUS mit 100 Sprachen und 40 G Token (vgl. Tiedemann/Nygaard 2004 und Tiedemann 2012), das multilinguale Parallelkorpus InterCorp mit 40 Sprachen und mit 1.5 G Token (vgl. Čermák/Rosen 2012; Rosen/Vavřín/Zasina 2019 sowie Káňa i. d. Bd.) und The European Parliament Proceedings Parallel Corpus Europarl mit 21 Sprachen und 0.6 G Token (Koehn 2005). Darüber hinaus gibt es eine Reihe von kleineren Parallelkorpora, die oft bilingual sind bzw. nur wenige Sprachen umfassen, dafür häufig aufgrund (teilweise) manueller Annotation detailliertere und genauere linguistische Information enthalten. Als Beispiel können The Stockholm MULtilingual TReebank SMULTRON (vgl. Volk et al. 2015), The CroCo Corpus (vgl. Hansen-Schirra/Neumann/Vela 2006) oder das Tschechisch-Englische Parallelkorpus CzEng (vgl. Bojar/Žabokrtský 2006) genannt werden.

Paralleldaten, wie sie in Parallelkorpora bereitgestellt werden, stellen sprachliche Einheiten (Wörter, Phrasen, Sätze) in zwei oder mehreren Sprachen dar, die Übersetzungsäquivalente voneinander sind (denen wiederum eine funktionale Äquivalenz zugrunde liegt) und als solche die gleiche (oder ähnliche) Bedeutung transportieren. Der große Vorteil ist auch, dass man diese sprachlichen Einheiten in den jeweiligen Quell- und Zielsprachen kontextbezogen und innerhalb der gleichen Texttypen bezogen auf genau die gleichen Themen, Zeiträume etc. betrachten kann. Aufgrund dieser Eigenschaften bieten Paralleldaten eine perfekte Grundlage für die Ermittlung funktionaler Äquivalenz zwischen sprachlichen Strukturen im sprachübergreifenden Kontext. In anderen Worten können sie als ein perfektes *tertium comparationis* verwendet werden (vgl. auch James 1980; Chesterman 1998). Darüber hinaus erlauben Paralleldaten Einblicke in sprachübergreifende Ähnlichkeiten und Divergenzen, die bei der Arbeit mit einsprachigen Korpora leicht übersehen werden könnten.

Diese Vorteile von Paralleldaten wurden früh in sprachübergreifender Forschung erkannt und wurden in zahlreichen Studien im Bereich der kontrastiven Linguistik (vgl. Altenberg und Granger 2002 oder Granger 2010, um nur einige Bei-

1 Als erstes (mehrsprachiges) Parallelkorpus gilt das Englisch-Norwegische Parallelkorpus Corpus ENPC, das 1994–1997 an der Universität Oslo entwickelt wurde und dessen zugrundeliegende Modell auch für weitere Sprachen, einschließlich Deutsch erfolgreich verwendet wurde.

spiele zu nennen), Sprachtypologie (vgl. u. a. Cysouw/Wälchli 2007 und andere Artikel jenes Heftes) sowie Übersetzungswissenschaften (vgl. z. B. Granger/Lerot/Petch-Tyson 2003) umgesetzt. Auch am IDS, und insbesondere im Projekt GDE kommen Parallelkorpora zum Einsatz (vgl. auch Trawiński/Schlotthauer/Bański i. d. Bd.). Hierzu kann die Studie zum Imperativ in den vier europäischen Sprachen Deutsch, Englisch, Polnisch und Tschechisch genannt werden, die zum Ziel die Validierung der *Agentivitätshypothese*[2] hatte (Trawiński 2016a, b). Wir gehen im Folgenden auf diese Studie etwas genauer ein, um zu demonstrieren, dass auch der Sprachvergleich mit Parallelkorpora trotz vieler Vorteile ernsthafte Defizite hat.

Als Datenquelle in Trawiński (2016a, b) wurde das Parallelkorpus InterCorp (Release 6) über die KonText-Schnittstelle verwendet. Die zugrundeliegende Datengrundlage setzte sich genauer aus den Texten zusammen, die in Tabelle 2 zusammengefasst sind.

Tab. 2: Übersicht über die Paralleltexte in Trawiński (2016a, b)

Titel (14)	Autoren	DE	EN	CZ	PL	RU
1984	George Orwell	114.009	***120.437***	98.302	98.088	93.349
Das chasarische Wörterbuch	Milorad Pavic	118.406	116.792	100.894	96.011	105.136
Der Alchimist	Paolo Coelho	44.058	47.786	36.912	38.826	38.786
Der Herr der Ringe: Die Rückkehr des Königs	John R. R. Tolkien	167.062	***158.991***	178.243	141.341	125.323
Der Herr der Ringe: Die zwei Türme	John R. R. Tolkien	188.149	***183.972***	154.331	166.885	149.982

2 Die Agentivitätshypothese besagt, dass Imperativmarker bei agentivischen Verben signifikant häufiger vorkommen als bei nicht-agentivischen Verben (vgl. auch Potsdam 1996; Jensen 2003).

Titel (14)	Autoren	DE	EN	CZ	PL	RU
Die Abenteuer des braven Soldaten Schwejk	Jaroslav Hašek	286.820	196.240	***247.340***	257.999	248.739
Die unerträgliche Leichtigkeit des Seins	Milan Kundera	98.240	99.464	***83.646***	84.691	87.443
Die Unsterblichkeit	Milan Kundera	120.755	121.278	***107.929***	100.344	106.540
Farm der Tiere	George Orwell	33.656	***34.434***	27.061	28.296	29.213
Meister und Margarita	Mikhail Bulgakov	157.036	162.666	137.530	151.312	***145.185***
Pippi Langstrumpf	Astrid Lindgren	30.960	31.127	25.362	27.368	27.515
Scherz	Milan Kundera	131.535	120.996	***110.228***	115.455	113.649
Wunderbare Reise des kleinen Nils Holgersson mit den Wildgänsen	Selma Lagerlöf	242.296	179.383	170.553	199.159	201.382
Der kleine Prinz	Antoine de Saint-Exupéry	18.477	21.199	15.519	14.764	16.071
***Größe* (*Tokens*)**		***1.751.459***	***1.594.765***	***1.493.850***	***1.520.539***	***1.488.313***

Aus diesen Texten wurden imperativische Wortformen extrahiert, die jeweils 50 häufigsten Lemmata pro Sprache identifiziert und die ausgewählten (sprachspezifischen) Lemmata auf abstrakte Konzepte basierend auf *FrameNet*-Frame-Index (Baker/Fillmore/Lowe 1998) abgebildet. Die quantitative Analyse der Daten hat die Agentivitätshypothese sprachübergreifend bestätigt.

Ein kurzer Blick auf Tabelle 2 macht jedoch deutlich, dass – bei hoher Vergleichbarkeit in Bezug auf Inhalt und Größe – die Studie auf einer verhältnismäßig

kleinen und undifferenzierten Datenbasis beruht (vgl. Tab. 1). Dabei gilt generell, dass je mehr Sprachen man zum Vergleich heranzieht, umso stärker die Anzahl und Differenziertheit der Paralleltexte abnimmt. Darüber hinaus weisen die Texte in Tabelle 2 eine starke Unausgewogenheit in Bezug auf Originaltexte (Fett-Markierung) und Übersetzungstexte auf. Das stellt insbesondere in Hinblick auf die Besonderheiten von Übersetzungstexten im Allgemeinen ein Problem dar.

Übersetzungstexte werden aufgrund ihrer speziellen Eigenschaften manchmal als dritter Code betrachtet, das heißt als eine besondere Art von Text, die sich sowohl von der Ausgangssprache als auch von der Zielsprache unterscheidet (vgl. Frawley 1984; Baker 1993). Laviosa (1998) spezifiziert die folgenden Merkmale von Übersetzungstexten: relativ geringer Anteil lexikalischer Wörter gegenüber Funktionswörtern, relativ hoher Anteil hochfrequenter Wörter gegenüber niedrigfrequenten Wörtern, häufige Wiederholung von häufigsten Wörtern und niedrige Varietät bei häufigsten Wörtern. Baker (1995) beobachtet weiterhin, dass Übersetzungen dazu tendieren, eine einfachere Sprache zu verwenden (Vereinfachung), Dinge zu verdeutlichen (Explikation) und typische Muster der Zielsprache übermäßig zu gebrauchen (Normalisierung). In Teich (2003) wird das Phänomen des *shining-through* neben der Normalisierung definiert und anhand von deutsch-englischen bzw. englisch-deutschen Korpora empirisch am Beispiel verschiedener grammatischer Konstruktionen (wie Passiv oder Relativsätze) untersucht. *Shining-through* tritt auf, wenn sich Übersetzungen stärker an der Ausgangssprache als an der Zielsprache orientieren. Die Normalisierung im Sinne von Teich (2003) liegt vor, wenn sich Übersetzungen stärker an der Zielsprache orientieren als es zu erwarten wäre.

Zusammenfassend lässt sich also sagen, dass Parallelkorpora hohe Vergleichbarkeit in Bezug auf Größe und Inhalt aufweisen, was für den Sprachvergleich von großer Bedeutung ist. Im Gegensatz dazu ist die Qualität des linguistischen Materials gering(er) im Vergleich zu monolingualen Korpora. Dieses Fazit lässt sich grafisch wie in Abbildung 2 veranschaulichen.

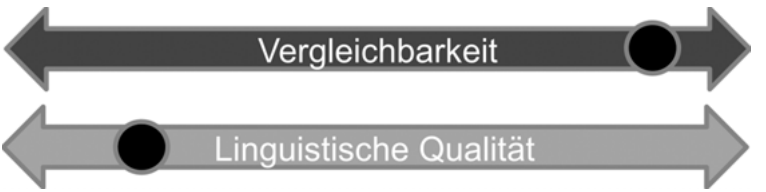

Abb. 2: Hohe Vergleichbarkeit und geringe sprachliche Qualität in parallelen Korpora

2.3 Vergleichbare Korpora

Wie wir oben festgestellt haben, eignen sich einsprachige und parallele Korpora allein nicht für feinkörnigere sprachübergreifende Forschung, weil es ihnen entweder an Vergleichbarkeit oder an linguistischer Qualität mangelt. Eine mögliche Abhilfe könnte die Verwendung einer Kombination aus parallelen und monolingualen Korpora sein, die jedoch für typische Anwendungsfälle kompliziert zu handhaben wäre. Es besteht daher ein klarer Bedarf an mehrsprachigen Korpora, die einerseits eine hohe Vergleichbarkeit in Bezug auf Inhalt und Größe gewährleisten und andererseits die Qualität der Originalsprache sicherstellen (siehe Abb. 3).

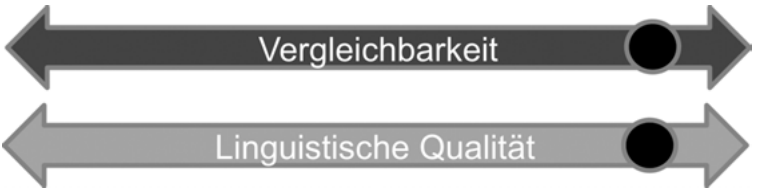

Abb. 3: Hohe Vergleichbarkeit und hohe sprachliche Qualität in einem idealen mehrsprachigen Korpus

Vergleichbare Korpora stellen eine interessante Option dar. Ein vergleichbares Korpus besteht aus zwei oder mehr einsprachigen Korpora, die hinsichtlich relevanter Eigenschaften wie Entstehungszeit, Medialität, Textsorte, Themenbereich usw. ähnlich aufgebaut sind und idealerweise nur Originaltexte enthalten. Ein frühes prominentes Beispiel für ein vergleichbares Korpus ist das International Corpus of English ICE (Greenbaum 1991), das zwölf Korpora verschiedener nationaler oder regionaler Varianten des Englischen mit einer kontrollierten, ähnlichen Zusammensetzung enthält. Im Jahr 2017 startete eine neue internationale Gemeinschaftsinitiative zum Aufbau des International Comparable Corpus ICC (Kirk/Čermáková 2017). Ziel dieser Initiative ist der Aufbau vieler kleiner Korpora mit kontrollierter Zusammensetzung nach dem Vorbild des ICE. Das primäre Ziel ist die Bereitstellung hochgradig vergleichbarer Datensätze für kontrastive Studien. Die derzeit beteiligten Sprachen sind Tschechisch, Finnisch, Französisch, Deutsch, Norwegisch, Polnisch, Slowakisch und Schwedisch. Das ICC ist ein laufendes Projekt und steht für die linguistische Forschung noch nicht zur Verfügung.

Gegenwärtig sind lediglich webbasierte vergleichbare Korpora verfügbar, wie zum Beispiel Aranea – Family of Comparable Gigaword Web Corpora (Benko 2014). Aranea umfasst Korpora von aktuell 20 Sprachen mit kontrollierter Größe von jeweils 1.2 G (die Maius-Ausgabe) und 120 M Token (die Minus-Ausgabe, eine 10%-ige Zufallsstichprobe von Maius). Die Ressource wurde mit frei verfügbaren Werkzeugen entwickelt und ist zum Beispiel über NoSketch Engine (Rychlý 2007) oder KonText (Machálek 2014, 2020) zugänglich (siehe Abb. 4).

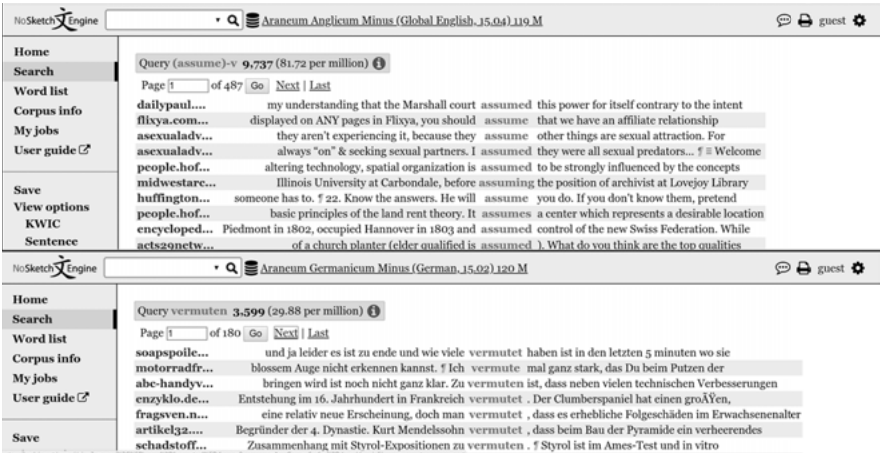

Abb. 4: Aranea im Einsatz mit der NoSketch Engine

Während die Größe der Aranea-Korpora kontrolliert ist, ist es ihre Komposition keineswegs. In Wirklichkeit kann sie auch gar nicht (leicht) kontrolliert werden, weil die erforderlichen Metadaten, wie Autorenschaft, Herausgeberschaft, Zeit und Ort der Veröffentlichung, Textart, Thema usw. für Texte aus dem Web meistens nicht vorhanden sind.

Vor diesem Hintergrund lässt sich festhalten, dass im Falle von webbasierten vergleichbaren Korpora weder das Kriterium der Vergleichbarkeit noch das Kriterium der sprachlichen Qualität ohne Weiteres erfüllt werden können (siehe Abb. 5).

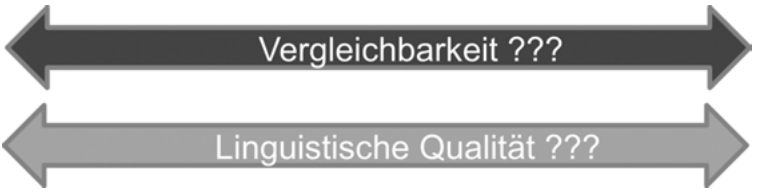

Abb. 5: Unkontrollierte Vergleichbarkeit und unkontrollierte linguistische Qualität in webbasierten Vergleichskorpora

3 Das Europäische Referenzkorpus EuReCo

Ziel der 2013 gegründeten offenen EuReCo-Initiative (Kupietz et al. 2017) ist es, den Mangel an qualitativ hochwertigen mehrsprachigen Vergleichskorpora zu beheben. Dabei geht es jedoch nicht darum, neue mehrsprachige Korpora aufzubauen, da dies zumindest ökonomisch schwer umzusetzen wäre, sondern auf den bestehenden einsprachigen Referenz- und nationalen Korpora aufzubauen und diese virtuell zu Tupeln vergleichbarer Korpora zusammenzuführen. Das bedeutet, dass die jeweiligen Korpora an ihren Standorten verbleiben und über eine gemeinsame Software-Infrastruktur vernetzt sind. Die virtuelle Verschmelzung ist dabei unerlässlich, da die Texte, aus denen nationale und Referenzkorpora bestehen, in der Regel durch Lizenzverträge, die zumindest das Kopieren ganzer Texte verbieten, an ihre Heimatinstitutionen gebunden sind. Dieses infrastrukturelle Problem wird derzeit durch die Korpusanalyseplattform KorAP (Bański et al. 2013) gelöst, die verteilte Indizes und die dynamische Definition virtueller Subkorpora unterstützt und die Korpusdaten über eine einheitliche Schnittstelle auch für die weitere linguistische Analyse zur Verfügung stellt. Der Aufbau vergleichbarer Korpora erfolgt auf der Basis von Textmetadaten in der Weise, dass im Idealfall der Nutzer oder die Nutzerin selbst dynamisch vergleichbare virtuelle Subkorpora definieren kann – perspektivisch durch einfache Befehle wie „Erstelle ein möglichst großes Korpuspaar mit identischer Zusammensetzung hinsichtlich Thema, Textsorte und Erscheinungsjahr“. Eine solche dynamische Definierbarkeit und Korrigierbarkeit mit der Möglichkeit der persistenten Speicherung ist wichtig, da das gegenüber einer normalen korpuslinguistischen Untersuchung zusätzliche Korpus und die zusätzliche Erforder-

nis der Vergleichbarkeit das Risiko von Artefakten, die allein durch die Korpuskomposition bedingt sind, zusätzlich steigern. Anders als in der monolingualen Korpuslinguistik muss nicht nur ein Korpus repräsentativ für eine intendierte Sprachdomäne in Bezug auf eine Forschungsfrage sein, sondern auch das zweite Korpus bzgl. seiner Sprache. Außerdem müssen beide Korpora vergleichbar sein. Das ohnehin schon hohe Risiko, Ergebnisse zu erhalten, die nichts über die intendierte Sprachdomäne aussagen, sondern nur durch eine schiefe Korpuszusammensetzung bedingt sind, ist dementsprechend höher, wenn mit vergleichbaren Korpora gearbeitet wird.

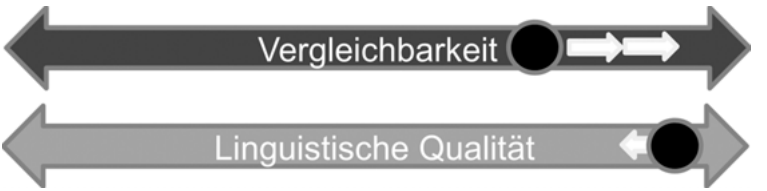

Abb. 6: Allmähliche Verbesserung der Vergleichbarkeit bei weitgehend gleichbleibend hoher sprachlicher Qualität durch iterative Verfeinerung der Metadatenzuordnungen und Vergleichbarkeitskriterien

Um Korpuszusammensetzungen anpassen zu können und damit eine schrittweise Verbesserung der Vergleichbarkeit zu ermöglichen, wenn der Verdacht besteht, dass die korpusbasierten Ergebnisse nur durch verzerrte Korpuszusammensetzungen bedingt sind, sollte der Konstruktionsprozess idealerweise iterativ angelegt sein (siehe Kupietz 2015, S. 64; Kupietz et al. 2020a).

Mit der Möglichkeit, Vergleichbarkeitskriterien und damit vergleichbare Korpuspaare dynamisch zu definieren und zu verfeinern, kann auch die Stabilität quantitativer Ergebnisse in Bezug auf unterschiedlich definierte Vergleichskorpora bewertet werden. Es ist jedoch zu beachten, dass die Flexibilität verschiedener vergleichbarer Korpusdefinitionen durch die Größe und Schichtung der zugrundeliegenden einsprachigen Korpora begrenzt ist und dass zusätzliche Vergleichbarkeitskriterien typischerweise die Größe der resultierenden vergleichbaren Korpuspaare reduzieren, so dass auch der Ansatz von EuReCo einen Kompromiss zwischen Vergleichbarkeit und Korpusgröße nicht vermeiden kann.

3.1 DRuKoLA: Das erste EuReCo-Projekt

Teile der EuReCo-Vision sind bereits im DRuKoLA-Projekt (2016–2018) umgesetzt worden.[3] Im Zentrum von DRuKoLA standen das Deutsche Referenzkorpus DeReKo, mit damals 42 Milliarden Wörtern (Kupietz et al. 2018) die größte Sammlung deutscher Texte, mit einem sogenannten Primordial-Sample-Design, das auch für die Definition verschiedener virtuell vergleichbarer Korpora im EuReCo-Kontext grundlegend ist, und das Referenzkorpus der rumänischen Gegenwartssprache CoRoLa (Tufiș et al. 2015; Barbu Mititelu/Tufiș/Irimia 2018), das fast eine Milliarde Wörter enthält und im Dezember 2017 öffentlich vorgestellt wurde und über verschiedene Schnittstellen, darunter KorAP, abgefragt werden kann (Cosma et al. 2016; Cristea et al. 2019).

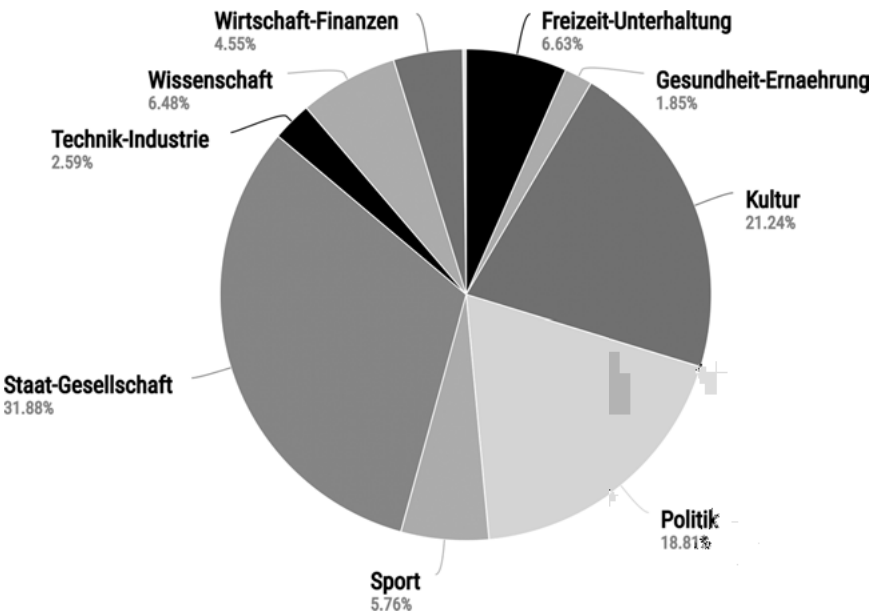

Abb. 7: Themenanteile (nach DeReKo's Top-Level-Domänen) im ersten vergleichbaren Korpus

3 DRuKoLA (2016–2018) wurde von der Alexander von Humboldt-Stiftung als Programm zur Vernetzung von Forschungsgruppen gefördert. Das Akronym verbindet zentrale Ziele des Projekts: Korpusentwicklung und kontrastive linguistische Analyse (*Sprachvergleich korpustechnologisch. Deutsch-Rumänisch*).

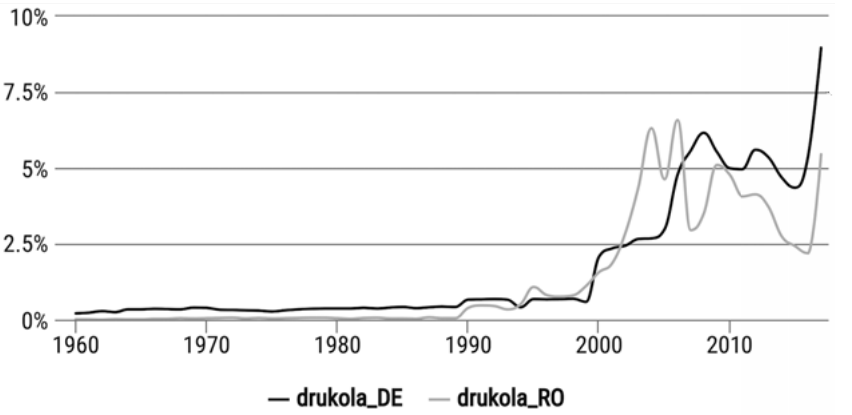

Abb. 8: Relative Größen pro Jahr der Veröffentlichung im Vergleich

Mit Abschluss des DRuKoLA-Projekts konnte ein erstes virtuelles Vergleichskorpus über KorAP öffentlich zugänglich gemacht werden, das vorerst ausschließlich auf einer Abbildung der Thementaxonomie von DeReKo auf die von CoRoLa basiert. Diese Zuordnung ist nicht perfekt, da sich die für die beiden Korpora verwendeten Klassifikationssysteme stark unterscheiden. Während für DeReKo eine Teilmenge der Open Directory (dmoz)-Taxonomie verwendet wurde (siehe Klosa et al. 2012), wurden für CoRoLa die Top-Level-Domänen der englischen Wikipedia und das System der Universal Decimal Classification (UDC) verwendet (siehe Gîfu et al. 2019). Mit Hilfe einer heuristischen Abbildung der Kategorien konnten aber fast 90% aller CoRoLa-Texte den DeReKo-Kategorien zugeordnet werden. Unter anderem, um diese Abbildung noch zu verbessern, plant das IDS, in Zukunft UDC- und Wikipedia-Domänen für DeReKo bereitzustellen.

Aufgrund des wesentlich größeren Umfangs von DeReKo und seiner hinreichend ähnlichen Streuung in Bezug auf die Themenbereiche war es möglich, das erste vergleichbare Korpus aufzubauen, indem nur eine Teilstichprobe von DeReKo definiert wurde, die die thematische Zusammensetzung des gesamten CoRoLa nachahmt (Abb. 7). Es ist anzumerken, dass wir für dieses erste vergleichbare deutsch-rumänische Korpus die Zusammensetzung im Hinblick auf Erscheinungsjahre und Textsorten nicht kontrolliert haben. Eine erste oberflächliche Untersuchung des deutschen Teils zeigt jedoch, dass zumindest ein breites Spektrum von Textsorten, wie Presseberichte, Leitartikel, Lexikonartikel, populärwissenschaftliche Artikel, Essays, Romane, Biografien, Lehrbücher, Tagebücher, Kinderbücher, Handbücher, politische Reden, Interviews, Gerichtsentscheidungen, Leserbriefe, Horoskope usw. (in absteigender Reihenfolge) im virtuellen

DeReKo-Subkorpus erfasst sind. Darüber hinaus ist auch, wie Abbildung 8 zeigt, die zeitliche Verteilung der Texte in den vergleichbaren Korpora recht ähnlich, obwohl diese, wie gesagt, nicht kontrolliert wurde. Weitere Untersuchungen werden zeigen, inwieweit auch die quantitative Verteilung der Textsorten angeglichen werden kann, ohne dass dabei das Vergleichskorpus zu klein wird.

3.2 Bilinguale Worteinbettungen als Grundlage für Vergleichbarkeitsmaße

Zu Beginn des DRuKoLA-Projekts war geplant, mit bilingualen Worteinbettungen (Zhou et al. 2013) für sprachvergleichende distributionell semantische Studien zu experimentieren und diese auch in Form von sogenannten semantischen Fingerabdrücken (*semantic fingerprints*) (Kutuzov et al. 2016; Saad/Langlois/Smaïli 2013) oder sogenannten Dokumenteinbettungen (Le/Mikolov 2014) als weiteres Maß für Textähnlichkeit über Sprachgrenzen hinweg zu verwenden. Für erste Vorstudien zum Projekt wurden dazu Worteinbettungen für DeReKo und CoRoLa berechnet und anhand eines zweisprachigen Lexikons[4] eine Matrix zur Transformation der CoRoLa-Embeddings in den DeReKo-Embedding-Raum berechnet. Im Projektverlauf zeigte sich jedoch nach einer Integration der transformierten CoRoLa-Embeddings in das DeReKoVecs-Tool[5] (Fankhauser/Kupietz 2019; Kupietz et al. 2018) und die IDS-Wortraumstation (Kupietz et al. 2020b), dass es zumindest nicht ohne größeren Aufwand möglich sein würde, ausreichend linguistisch plausible und interessante deutsch-rumänische Worteinbettungen zu erhalten. Aus diesem Grund wurden die Versuche mit dieser Methodik nicht weiter systematisch fortgeführt. Ein weiterer Grund für diese Entscheidung war, dass DeReKo und CoRoLa mit reichhaltigen Metadaten versehen waren, so dass keine Notwendigkeit bestand, auf einbettungsbasierte Vergleichbarkeitsmaße zurückzugreifen.

4 Zur Erzeugung des Wörterbuchs wurden die OPUS-Korpora und -Tools verwendet (Tiedemann 2012).

5 http://corpora.ids-mannheim.de/openlab/derekovecs (Stand: 29.10.2020)

3.3 DeutUng

Im zweiten EuReCo-Pilotprojekt, DeutUng,[6] wurde damit begonnen, das Ungarische Nationalkorpus HNC (Váradi 2002; Oravecz/Váradi/Sass 2014) in EuReCo zu integrieren. Der aktuelle Stand der DeutUng ist, dass ein Konverter für das HNC-Format in das Eingabeformat von KorAP entwickelt wurde und eine erste HNC-Stichprobe über KorAP zur Verfügung steht, die bereits für erste Pilotstudien genutzt wird (siehe Absch. 3.2). Das gesamte HNC sowie entsprechend größere vergleichbare deutsch-ungarische Korpora sollen bis Ende 2020 über KorAP nutzbar sein.

4 Zugang zu vergleichbaren Korpora mit KorAP

Wie bereits erwähnt, ist die aktuelle technische Grundlage für EuReCo die Korpusabfrage- und Analyseplattform KorAP[7] (Bański et al. 2013; Diewald et al. 2016; Diewald et al. i. d. Bd.), die derzeit am IDS entwickelt wird und seit Mai 2017 als öffentliche Beta-Version verfügbar ist. KorAP ist der designierte Nachfolger des Korpussuch- und Verwaltungssystems COSMAS II[8] als Hauptzugang zu DeReKo. KorAP ist für Korpora in verschiedenen Sprachen mit unterschiedlichen Annotationen anpassbar. Sie unterstützt auch mehrere Korpusabfragesprachen (z. B. Poliqarp, COSMAS II QL, AnnisQL), so dass Nutzer aus unterschiedlichen fachlichen und sprachspezifischen Communities optimal unterstützt werden. Darüber hinaus stellt KorAP auch Client-Bibliotheken für die Programmiersprachen R und Python zur Verfügung (Kupietz/Diewald/Margaretha 2020). Für vergleichbare Korpora im EuReCo-Szenario bietet KorAP einige wesentliche Merkmale, insbesondere

- die Fähigkeit, Korpora zu verwalten, die sich physisch an verschiedenen Orten befinden, so dass typische Lizenzbeschränkungen leicht einzuhalten sind (Kupietz et al. 2014);
- die Möglichkeit, virtuelle Subkorpora auf der Grundlage von Texteigenschaften dynamisch zu erstellen und diese persistent zu verwalten, um beispielsweise Wiederverwendbarkeit und Reproduzierbarkeit garantieren zu können;

---

6 DeutUng (2017–2020) ist ein Kooperationsprojekt zwischen dem IDS und der Universität Szeged mit dem Forschungsinstitut für Linguistik der Ungarischen Akademie der Wissenschaften als assoziiertem Partner. DeutUng wird wie das DRuKoLA-Projekt von der Alexander von Humboldt-Stiftung gefördert.

7 https://korap.ids-mannheim.de/ (Stand: 29.10.2020)

8 https://cosmas2.ids-mannheim.de/ (Stand: 29.10.2020)

– die Möglichkeit, virtuelle Korpora mit Hilfe der Programmierschnittstellen dynamisch und explorativ zu definieren und zu optimieren und damit den oben dargestellten iterativen Konstruktionsprozess vergleichbarer Korpora automatisieren zu können;
– die Möglichkeit quantitative Sprachvergleichsstudien auch mit variablen Parametern mit Hilfe der Programmierschnittstellen einfach und nachvollziehbar umzusetzen.

4.1 Zugang zum deutsch-rumänischen Vergleichskorpus

Für CoRoLa wurde ein in Größe und Zusammensetzung vergleichbares Teilkorpus von DeReKo auf der Grundlage von Textmetadaten zusammengestellt. Dieses Subkorpus ist als persistentes virtuelles Korpus (VC) in KorAP gespeichert und kann (optional als Teil eines komplexeren VC) referenziert werden,[9] um die Suche und Analyse auf alle Dokumente im vergleichbaren Korpus einzuschränken. Das deutsch-rumänische Vergleichskorpus besteht derzeit aus mehr als 3 Millionen Dokumenten, die 940 Millionen Worttoken umfassen. Obwohl sich Metadaten und Annotationen unterscheiden, können beide Korpora in KorAP auf vergleichbare Weise durchsucht werden. Abbildung 9 zeigt zum Beispiel eine Abfrage nach postnominalen Adjektivfolgen, die in beiden Korpora durchgeführt wurde, wobei zwar die Trefferverteilung, wie erwartet, ein häufigeres postnominales Muster im Rumänischen bestätigt, im Detail aber Korrelationen z. B. mit der Textsorte aufweist, und damit neue, noch zu überprüfende Hypothesen aufwirft.

Eine eingehende Studie kann dann diese unterschiedlichen Muster bezüglich der Adjektivpositionen in beiden Korpora vergleichen, indem die Abfragen verfeinert werden, um sprachspezifische Annotationen zu erkennen (vgl. Cornilescu/Cosma 2019).

9 Der Referenzbezeichner lautet ‚drukola.20180909.1b_words'.

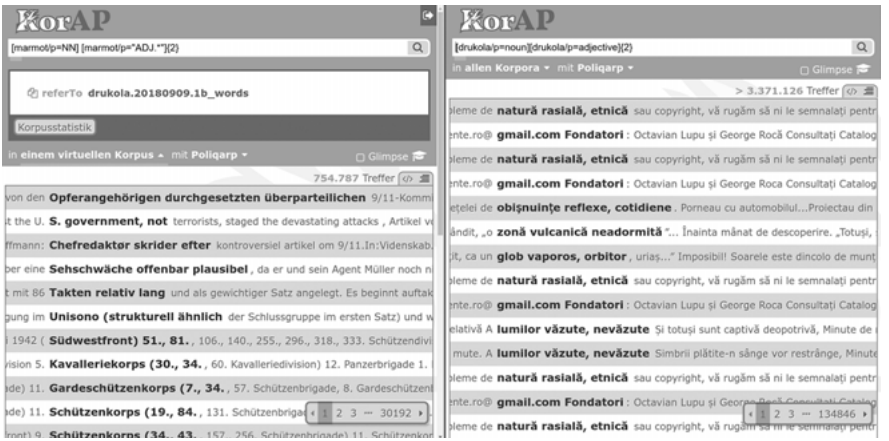

Abb. 9: Suche nach einem Nomen gefolgt von zwei Adjektiven im ersten deutsch-rumänischen Vergleichskorpus, ausgedrückt in der Anfragesprache Poliqarp QL bezugnehmend auf verschiedene zugrunde liegende Annotationen

4.2 Zugang zum deutsch-ungarischen Vergleichskorpus

Im Rahmen des DeutUng-Projekts wurden erste Teile des HNC in EuReCo integriert und kleine deutsch-ungarische Vergleichskorpora stehen zur Abfrage bei KorAP zur Verfügung.

Eine der im DeutUng-Projekt behandelten linguistischen Forschungsfragen ist die Verwendung von Korrelaten bei Komplementsätzen (Hartmann et al. 2017). Im Ungarischen ist das Korrelat *azt* in Strukturen mit assertiven Verben (wie z. B. *sagen*) möglich, nicht aber in Strukturen mit Faktiverben (wie *bedauern*). Im Deutschen ist genau das Gegenteil der Fall: Das Korrelat *es* kann in komplexen Sätzen mit Faktivverben verwendet werden, nicht aber mit assertiven Verben (siehe Molnár 2015, S. 211 f.; Kupietz et al. 2020a). Molnár (2015) weist jedoch darauf hin, dass in bestimmten Kontexten bzw. unter bestimmten Umständen (z. B. Fokus), das ungarische Korrelat *azt* auch mit faktiven Prädikaten möglich zu sein scheint. Auch die Verwendung des deutschen Korrelats *es* ergibt kein klares Bild, wenn unterschiedliche (z. B. informationstrukturelle) Bedingungen berücksichtigt werden. Das Ziel der im Rahmen des DeutUng-Projekts vorgesehenen kontrastiven Untersuchung ist es, die Faktoren zu identifizieren, die die Verwendung der Korrelate in den beiden Sprachen determinieren.

Abbildung 10 zeigt einen Ausschnitt von Suchergebnissen, die die Suche nach Korrelaten mit bestimmten faktiven bzw. assertiven Verben in DeReKo und HNC mittels KorAP ergibt (siehe auch Kupietz et al. 2020a).

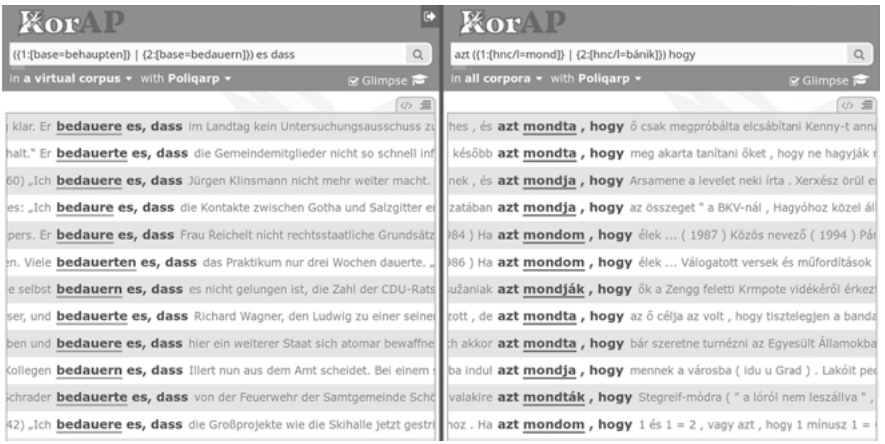

Abb. 10: Suche in DeReKo und HNC nach Korrelaten mit faktiven bzw. assertiven Verben (Die unterschiedliche Hervorhebung der Verbtypen deutet auf eine umgekehrte Verwendung des Musters hin.)

5 Schlussfolgerungen und Ausblick

Wir haben gezeigt, wie die EuReCo-Initiative den derzeitigen Mangel an mehrsprachigen Korpora beheben kann und dabei sowohl das Kriterium einer hohen sprachlichen Qualität, einschließlich Größe und Vielfalt, als auch das Kriterium der Vergleichbarkeit erfüllen kann. Wir zeigten auch, wie dies auf wirtschaftlich und rechtlich realistische Weise geschehen kann, und zwar indem wir auf bestehenden Korpora aufbauen, diese wiederverwenden und sie virtuell mit Hilfe der Korpusabfrageplattform KorAP zusammenführen. Darüber hinaus haben wir den Ansatz von EuReCo skizziert, wie die komplexe und fehleranfällige Definition von Vergleichbarkeit durch sukzessive Anpassung der Vergleichbarkeitskriterien angegangen werden kann. Schließlich haben wir gezeigt, wie die EuReCo-Ansätze in den Pilotprojekten DRuKoLA und DeutUng in ersten vergleichenden deutsch-rumänischen bzw. deutsch-ungarischen Studien umgesetzt werden konnten.

Als nächste Schritte stehen neben der Verbesserung von DeReKos Themengebietsklassifikation und der kontinuierlichen Weiterentwicklung von KorAP, z. B. hinsichtlich verteilt berechneter Aggregationsfunktionen und einer Oberfläche zum Management virtueller (vergleichbarer) Korpora, besonders die Erweiterung von EuReCo um neue Sprachen an.

Literatur

Altenberg, Bengt/Granger, Sylviane (2002): Lexis in contrast. Corpus-based approaches. (= Studies in Corpus Linguistics 7). Amsterdam/Philadelphia: Benjamins.

Augustin, Hagen (2018): Verschmelzung von Präposition und Artikel. Eine kontrastive Analyse zum Deutschen und Italienischen. (= Konvergenz und Divergenz 6). Berlin/Boston: De Gruyter.

Baker, Mona (1993): Corpus linguistics and translation studies – Implications and applications. In Baker, Mona/Francis, Gill/Tognini-Bonelli, Elena (Hg.): Text and Technology: In honour of John Sinclair. Amsterdam/Philadelphia: Benjamins, S. 233–250.

Baker, Mona (1995): Corpora in translation studies. An overview and some suggestions for future research. In: Target 7, 2, S. 223–243.

Baker, Collin F./Fillmore, Charles J./Lowe, John B. (1998): The Berkeley FrameNet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Montreal: ACL.

Bański, Piotr/Bingel, Joachim/Diewald, Nils/Frick, Elena/Hanl, Michael/Kupietz, Marc/Pęzik, Piotr/Schnober, Carsten/Witt, Andreas: (2013): KorAP: the new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt/Uszkoreit, Hans (Hg.): Human language technologies as a challenge for computer science and linguistics. Proceedings of the 6th Language and Technology, Dezember 2013, Posen. Mannheim: Institut für Deutsche Sprache, S. 586–587.

Barbu Mititelu, Verginica/Tufiș, Dan/Irimia, Elena (2018): The reference corpus of the contemporary romanian language (CoRoLa). In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu/Goggi, Sara/Mazo, Hélène (Hg.): Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki/Paris: ELRA, S. 1178–1185.

Benko, Vladimír (2014): Aranea: Yet another family of (comparable) web corpora. In: Sjka, Petr/Horák, Ales/Kopeček, Ivan/Pala, Karel (Hg.): Text, speech and dialogue. 17th International Conference (TSD 2014), September 2014, Brünn. (= Lecture notes in computer science 8655). Cham/Heidelberg/New York: Springer, S. 247–256.

Bojar, Ondrej/Žabokrtský, Zdenek (2006): CzEng: Czech-English parallel corpus, release version 0.5. In: Prague Bulletin of Mathematical Linguistics 86, S. 59–62.

Borin, Lars/Forsberg, Markus/Roxendal, Johan (2012): Korp – the corpus infrastructure of Språkbanken. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Doğan, Mehmet/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios (Hg): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul/Paris: ELRA, S. 474–478.

Brandt, Patrick/Trawiński, Beata/Wöllstein, Angelika (2016): (Anti-)Control in German: evidence from comparative, corpus- and psycholinguistic studies. In: Reich, Ingo/Speyer, Augustin (Hg.): Co- and subordination in German and other languages. (= Linguistische Berichte – Sonderhefte 21). Hamburg: Buske, S. 77–98.

Čermák, Frantisek/Rosen, Alexandr (2012): The case of InterCorp, a multilingual parallel corpus. In: International journal of corpus linguistics 17, 3, S. 411–427.

Chesterman, Andrew (1998): Contrastive functional analysis. (= Pragmatics & beyond. New Series 47). Amsterdam/Philadelphia: Benjamins.

Cornilescu, Alexandra/Cosma, Ruxandra (2019): Linearization of attributive adjectives in Romanian and German. In: Revue roumaine de linguistique 3, S. 307–322.

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. Proceedings of LREC 2016. Portorož/Paris: ELRA, S. 28–32.

Cristea, Dan/Diewald, Nils/Haja, Gabriela/Mărănduc, Cătălina/Barbu Mititelu, Verginica/Onofrei, Mihaela (2019): How to find a shining needle in the haystack. Querying CoRoLa: solutions and perspectives. In: Revue roumaine de linguistique 64, 3, S. 279–292.

Cysouw, Michael/Wälchli, Bernhard (2007): Parallel texts: using translational equivalents in linguistic typology. In: STUF – Sprachtypologie und Universalienforschung 60, 2, S. 95–99.

Diewald, Nils/Hanl, Michael/Margaretha, Eliza/Bingel, Joachim/Kupietz, Marc/Bański, Piotr/Witt, Andreas (2016): KorAP Architecture – Diving in the deep sea of corpus data. In: Calzolari, Nicolette/Choukri, Khalid/Declerck, Thierry/Goggi, Sara/Grobelnik, Marko/Maegaard, Bente/Mariani, Joseph/Mazo, Helene/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 10th international conference on language resources and evaluation (LREC 2016). Portorož/Paris: ELRA, S. 4353–4360.

Fankhauser, Peter/Kupietz, Marc (2019): Analyzing domain specific word embeddings for a large corpus of contemporary German. International corpus linguistics conference, Cardiff, Juli 2019. Mannheim: Leibniz-Institut für Deutsche Sprache. Internet: https://doi.org/10.14618/ids-pub-9117 .

Frawley, William (1992): Linguistic semantics. Hillsdale: Erlbaum.

Gîfu, Daniela/Moruz, Alex/Bolea, Cecilia/Bibiri, Anca/Mitrofan, Maria (2019): The methodology of building CoRoLa. In: Revue roumaine de linguistique 64, 3, S. 241–253.

Givón, Talmy (1990): Syntax: a functional-typological introduction. Bd. 2. Amsterdam/Philadelphia: Benjamins.

Granger, Sylviane (2010): Comparable and translation corpora in cross-linguistic research. Design, analysis and applications. In: Journal of Shanghai Jiaotong University 2, S. 14–21.

Granger, Sylviane/Lerot, Jacques/Petch-Tyson, Stephanie (2003): Corpus-based approaches to contrastive linguistics and translation studies. Amsterdam/New York: Rodopi.

Greenbaum, Sidney (1991): The development of the international corpus of English. In: Aijmer, Karin/Altenberg, Bengt (Hg.): English corpus linguistics: Studies in honour of Jan Svartvik. London: Longman, S. 83–92.

Hansen-Schirra, Silvia/Neumann, Stella/Vela, Mihaela (2006): Multi-dimensional annotation and alignment in an English-German translation corpus. In: Proceedings of the 5th workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing, April 2006, Trient. Stroudsburg: ACL, S. 35–42.

Hartmann, Jutta M./Mucha, Anne/Trawiński, Beata/Wöllstein, Angelika (2018): Selectional preferences for (non-)finite structures as indicators of control relations: A cross-Germanic corpus study. Vortrag, Grammar and Corpora 2018, November 2018, Universität Paris-Diderot, Paris.

Hartmann, Jutta M./Schlotthauer, Susan/Trawiński, Beata/Wöllstein, Angelika (2017): Sprachvergleich: Einblicke in die aktuelle kontrastive Forschung am IDS: Nominal- und Verbgrammatik. Vortrag, Kick-off zum Projekt DeutUng, Oktober 2017, Universität Szeged.

James, Carl (1980): Contrastive Analysis. (= Applied Linguistics and Language Study Series). London: Longman.

Jensen, Britta (2003). Syntax and semantics of imperative subjects. In: Nordlyd 31, 1, S. 150–164.

Kirk, John/Čermáková, Anna (2017): From ICE to ICC: The new International Comparable Corpus. In: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (Hg.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, July 2017. Mannheim: Institut für Deutsche Sprache, S. 7–12.

Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. (= Lexicographica. International Annual for Lexicography 28). Berlin/Boston: De Gruyter, S. 71–97.

Koehn, Philipp (2005): Europarl: A parallel corpus for statistical machine translation. In: The 10th Machine Translation Summit (MT Summit). Proceedings of Conference, September 2005, Phuket, S. 79–86. Internet: https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf (Stand: 28.10.2020).

Kupietz, Marc (2015): Constructing a Corpus. In: Durkin, Philip (Hg.): The Oxford Handbook of Lexicography. Oxford: Oxford University Press, S. 62–75.

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, Helene/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020). Marseille/Paris: ELRA, S. 7015–7021.

Kupietz, Marc/Lüngen, Harald/Bański, Piotr/Belica, Cyril (2014): Maximizing the potential of very large corpora. In: Kupietz, Marc/Biber, Hanno/Lüngen, Harald/Bański, Piotr/Breiteneder Evelyn/Mörth, Karlheinz/Witt, Andreas/Takhsha, Jani (Hg.): Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014) – Workshop challenges in the management of large corpora (CMLC2). Reykjavik/Paris: ELRA, S. 1–6.

Kupietz, Marc/Lüngen, Harald/Kamocki, Pawel/Witt, Andreas (2018): The German reference corpus DeReKo: New developments – new opportunities. In: Calzolari/Choukri/Cieri/Declerck/Hasida/Isahara/Maegaard/Mariani/Moreno/Odijk/Piperidis/Tokunaga/Goggi/Mazo (Hg.), S. 3586–3591.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza/Bodmer, Franck/Stallkamp, Helge/Harders, Peter (2020b): Recherche in Social-Media-Korpora mit KorAP. In: Marx, Konstanze/Lobin, Henning/Schmidt, Axel (Hg.): Deutsch in Sozialen Medien. Interaktiv, multimodal, vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin/Boston: De Gruyter, S. 373–378.

Kupietz, Marc/Witt, Andreas/Bański, Piotr/Tufiş, Dan/Cristea, Dan/Váradi, Tamás (2017): EuReCo – Joining forces for a european reference corpus as a sustainable base for cross-linguistic research. In: Bański/Kupietz/Lüngen/Rayson/Biber/Breiteneder/Clematide/Mariani/Stevenson/Sick (Hg.), S. 15–19.

Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020a): Recent developments in the European Reference Corpus (EuReCo). In: Granger, Sylviane/Lefer, Marie-Aude (Hg.): Translating and comparing languages: Corpus-based insights. (= Corpora and Language in Use 6). Louvain-la-Neuve: Presses universitaires de Louvain, S. 257–273.

Kutuzov, Andrej/Kopotev, Mikhail/Sviridenko, Tatyana/Ivanova, Lyubov (2016): Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints. In: Rapp, Reinhard/Zweigenbaum, Pierre/Sharoff, Serge (Hg.): Proceedings of the 9th Workshop on Building and Using Comparable Corpora (BUCC 2016). Portorož: ELRA, S. 3–10. Internet: https://comparable.limsi.fr/bucc2016/pdf/BUCC02.pdf (Stand: 20.10.2020).

Laviosa, Sara (1998): Core patterns of lexical use in a comparable corpus of english narrative prose. In: Meta, 43, 4, S. 557–570.

Le, Quoc/Mikolov, Tomas (2014): Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning (ICML 2014). Peking: PMLR, S. 1188–1196.

Machálek, Tomas (2014): KonText – Corpus query interface. FF UK, Prag. http://kontext.korpus.cz/ (Stand: 16.10.2020).

Machálek, Tomas (2020): KonText: Advanced and flexible corpus query interface. In: Proceedings of the 12th conference on language resources and evaluation (LREC 2020), Marseille, Mai 2020. Marseille: ELRA, S. 7003–7008.

Molnár, Valeria (2015): The predicationality hypothesis. The case of Hungarian and German. In: Kiss, Katalin É./Surányi, Balazs/Dékány, Eva (Hg.): Approaches to Hungarian. Bd. 14. Papers from the 2013 Piliscsaba Conference. Amsterdam/Philadelphia: Benjamins, S. 209–244.

Oravecz, Csaba/Váradi, Tamas/Sass, Balint (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: ELRA, S. 1719–1723.

Potsdam, Eric (1996): Syntactic issues in the English imperative. Doktorarbeit. University of California Santa Cruz. New York/London: Garland.

Rapp, Irene/Laptieva, Ekatarina/Koplenig, Alexander/Engelberg, Stefan (2017): Lexikalisch-semantische Passung und argumentstrukturelle Trägheit – eine korpusbasierte Analyse zur Alternation zwischen *dass*-Sätzen und *zu*-Infinitiven in Objektfunktion. In: Deutsche Sprache 45, S. 193–221.

Rosen, Alexandr/Vavřín, Martin/Zasina, Adrian J. (2019): The InterCorp Corpus – Czech1), 12. Version. Prag: Institute of the Czech National Corpus/Charles University. Internet: https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze12#fn__1 (Stand: 20.10.2020).

Rychlý, Pavel (2007): Manatee/Bonito – A modular corpus manager. In: Sojka, Petr/Horák, Aleš (Hg.): First workshop on Recent Advances in Slavonic Natural Language Processing (Raslan). Brünn: Masaryk University, S. 65–70.

Saad, Motaz/Langlois, David/Smaïli, Kameli (2013): Extracting comparable articles from Wikipedia and measuring their comparabilities. In: Procedia – Social and behavioral sciences 95, S. 40–47.

Taborek, Janus (2018): Korpusbasiertes kontrastives Beschreibungsmodell für Funktionsverbgefüge. In: Schmale, Günter (Hg.): Lexematische und polylexematische Einheiten des Deutschen. (= Eurogermanistik 35). Tübingen: Stauffenburg, S. 135–154.

Teich, Elke (2003): Cross-Linguistic variation in system and text: A methodology for the investigation of translations and comparable texts. (= Text, Translation, Computational Processing 5). Berlin/Boston: De Gruyter Mouton.

Tiedemann, Jörg (2012): Parallel data, tools and interfaces in OPUS. In: Calzolari/Choukri/Declerck/Doğan/Maegaard/Mariani/Odijk/Piperidis (Hg.), S. 2214–2218.

Tiedemann, Jörg/Nygaard, Lars (2004): The OPUS corpus – parallel & free. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lissabon: ELRA, S. 1183–1186.

Trawiński, Beata (2016a): Messung der Distanz zwischen grammatischen Kategorien im sprachübergreifenden Kontext. In: Averina, Anna V. (Hg.): Grammatitscheskije kategorii v kontrastivnom aspektje. Sbornik nautschnych statjej no materialam mjeschdunarodnoj konfjerjentschii, Moskva, 11–14 maja 2016 [Grammatische Kategorien aus kontrastiver Sicht. Sammelband zur internationalen Konferenz, Moskau, 11.–14.05.2016], Bd. 1. Moskau: Moskauer Städtische Universität, S. 116–120.

Trawiński, Beata (2016b): Zur Vergleichbarkeit grammatischer Kategorien. Ein vektorbasierter Ansatz. In: Zhu, Jianhua/Zhao, Jin/Szurawitzki, Michael (Hg.): Akten des XIII. Internationalen Germanistenkongresses Shanghai 2015. Germanistik zwischen Tradition und Innovation. Bd. 2. Angewandte Sprachforschung. (= Publikationen der Internationalen Vereinigung für Germanistik (IVG) 21). Frankfurt a. M.: Lang.

Tufiş, Dan/Barbu Mititelu, Verginica/Irimia, Elena/Dumitrescu, Ştefan Daniel/Boroş, Tiberiu/Teodorescu, Nicolai Horai/Cristea, Dan/Scutelnicu, Andrei/Bolea, Cecilia/Moruz, Alex/Pistol, Laura (2015): CoRoLa starts blooming – An update on the reference corpus of contemporary Romanian language. In: Bański, Piotr/Biber, Hanno/Breiteneder, Evelyn/Kupietz, Marc/Lüngen, Harald/Witt, Andreas (Hg.): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC 3), Lancaster, July 2015. Mannheim: Institut für Deutsche Sprache, S. 5–10.

Van Noord, Gertjan/Schuurman Ineke/Vandeghinste, Vincent (2006): Syntactic annotation of large corpora in STEVIN. In: Calzolari, Nicoletta/Choukri, Khalid/Gangemi, Aldo/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Tapias, Daniel (Hg): Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genua/Paris: ELRA, S. 1811–1814.

Van Noord, Gertjan/Bouma, Gosse/van Eynde, Frank/de Kok, Daniel/van der Linde, Jelmer/Schuurman, Ineke/Tjong Kim Sang, Erik/Vandeghinste, Vincent (2013): Large scale syntactic annotation of written Dutch: Lassy. In: Spyns, Peter/Odijk, Jan (Hg.): Essential speech and language technology for Dutch. Results by the STEVIN programme. Heidelberg/New York/Dordrecht/London: Springer, S. 147–163.

Váradi, Tamas (2002): The Hungarian national corpus. In: Rodríguez, Manuel/Araujo, Carmen (Hg.): In Proceedings of the 3th International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas/Paris: ELRA, S. 385–389.

Volk, Martin/Göhring, Anne/Rios, Annette/Marek, Torsten/Samuelsson, Yvonne (2015): SMULTRON (4. Version) — The Stockholm MULtilingual parallel TReebank. Zürich: Institute of Computational Linguistics, Universität Zürich.

Wöllstein, Angelika (2015): Grammatik – explorativ. Hypothesengeleitete und -generierende Exploration variierender Satzkomplementationsmuster im standardnahen Deutsch. In: Eichinger, Ludwig M. (Hg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. (= Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/Boston: De Gruyter, S. 93–120.

Zhou, Will Y./Socher, Richard/Cer, Daniel/Manning, Christopher D. (2013): Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, October 2013. Seattle: ACL, S. 1393–1398.