Originalveröffentlichung in: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (Hrsg.):
Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch.
– Berlin [u.a.]: de Gruyter, 2021. S. 189-208. (Jahrbuch des Instituts für Deutsche Sprache 2020)
DOI: *https://doi.org/10.1515/9783110731514-011*

Erhard Hinrichs (Mannheim/Tübingen)

Multilinguale Sprachressourcen für die linguistische Forschung

Abstract: Sprachressourcen in digitaler Form liegen für ein immer breiteres Spektrum von Einzelsprachen vor. Linguistisch annotierte Korpora ermöglichen es, gezielt nach linguistischen Mustern auf der Wort-, Phrasen-, und Satzebene zu suchen und in quantitativer und qualitativer Hinsicht auszuwerten. In diesem Beitrag illustriere ich anhand von ausgewählten Beispielen den Mehrwert, den annotierte Textkorpora für die sprachwissenschaftliche Forschung bieten können. Viele der vorgestellten Sprachressourcen werden im Rahmen der CLARIN-Infrastruktur nachhaltig zur Verfügung gestellt. Die Korpora sind entweder durch Suchportale recherchierbar oder werden per Download zur Verfügung gestellt.

1 Einleitung

Die Jahrestagung 2020 des Leibniz-Instituts für Deutsche Sprache (IDS) hat den thematischen Schwerpunkt *Deutsch in Europa*. Die Tagung widmet sich u. a. der Frage, wie digitale Forschungsdaten für vergleichende sprachwissenschaftliche Untersuchungen genutzt werden können. Ich möchte auf diese Frage exemplarisch eine Antwort geben anhand von annotierten Textkorpora und anhand von sogenannten *Baumbanken*, deren Annotation nicht nur die Wortebene, sondern auch syntaktische Annotationen umfasst. In Abschnitt 2 werde ich ausgewählte Textkorpora für das Deutsche und für sprachvergleichende Untersuchungen vorstellen und die Nutzbarkeit dieser Sprachressourcen für die linguistische Forschung anhand von Fallbeispielen illustrieren. Die Beispiele beziehen sich sowohl auf die Wort- als auch auf die Satzebene. Abschnitt 3 gibt einen Überblick über Baumbanken des Deutschen und über die Initiative der *Universal Dependency Treebanks*, die Baumbanken mit dependenzgrammatischen Annotationen für ein breites Spektrum von Einzelsprachen zur Verfügung stellt. Abschnitt 4 stellt die Forschungsdateninfrastruktur CLARIN und die CLARIN Webanwendung WebLicht vor, die es erlaubt, Textkorpora mit computerlinguistischen Methoden automatisch mit linguistischen

https://doi.org/10.1515/9783110731514-011

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-103285

Annoationen anzureichern. Mein Beitrag endet mit Abschnitt 5, der einen Überblick über weitere Bereiche der Sprachwissenschaft gibt, in denen digitale Sprachressourcen eine zentrale Rolle spielen.

2 Ausgewählte Textkorpora für das Deutsche und für sprachvergleichende Untersuchungen

Mit dem Deutschen Referenzkorpus DeReKo (Kupietz et al. 2010) stellt das IDS die weltweit größte Sammlung deutschsprachiger elektronischer Korpora zur Verfügung. In der Version vom Januar 2020 umfasst DeReKo 46,9 Milliarden Wörter. Die Webseiten des IDS geben Auskunft über den nicht-linearen Zuwachs an Daten in den letzten Jahren: Der Datenumfang von DeReKo hat sich zwischen 1992 und 2015 von ca. 28 Millionen auf 28 Milliarden Textwörter erhöht. DeReKo kann mit Hilfe der on-line Werkzeuge COSMAS II (Bodmer 2005) und KorAP (Diewald et al. 2016) online durchsucht werden. Diese beiden Werkzeuge unterstützen neben Suchanfragen auch die Kompilierung von virtuellen Korpora. Solche virtuellen Teilkorpora können entweder nach Gesichtspunkten der Repräsentativität für eine bestimmte linguistische Fragestellung oder bewusst auf bestimmte Textsorten beschränkt werden.

Das DeReKo Korpus beinhaltet Texte, die bis ins 19. Jahrhundert zurückreichen. Daher können mit Hilfe von COSMAS II und KorAP Neologismen und Veränderungen in der Häufigkeit von Wörtern nachverfolgt werden. So ergibt eine Suchanfrage für das Wort *Energiewende* einen rasanten Anstieg von einem Treffer für die Dekade 1970–1979 auf 43.672 für den Zeitraum 2010–2019. Vor 1970 gibt es überhaupt keine Treffer. Da der Datenumfang von DeReKo für die in Frage stehenden Dekaden starken Schwankungen unterliegt, ist nicht nur die absolute Anzahl an Belegstellen relevant, sondern vor allem die relative Häufigkeit im Verhältnis zur Größe des Korpus. Abbildung 1 zeigt daher die Häufigkeit des Suchbegriffs pro 1 Million Wörter an und belegt, dass die relative Häufigkeit zwischen 2010 und 2019 überproportional angestiegen ist.

Neben lexikalischen Untersuchungen zu Einzelwörtern können mit COSMAS-II auch Kollokationsanalysen durchgeführt werden. Abbildung 2 zeigt die zehn hochrangigsten Kollokate für alle Wortformen des Adjektivs *schwarz*.

Abb. 1: Ergebnisse der COSMAS-II Suchanfrage *Energiewende* im Korpus DeReKo 2020-I

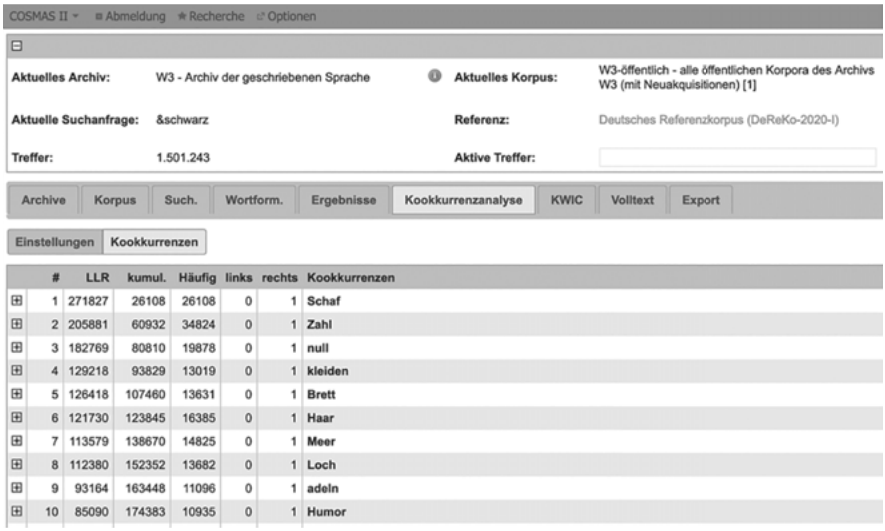
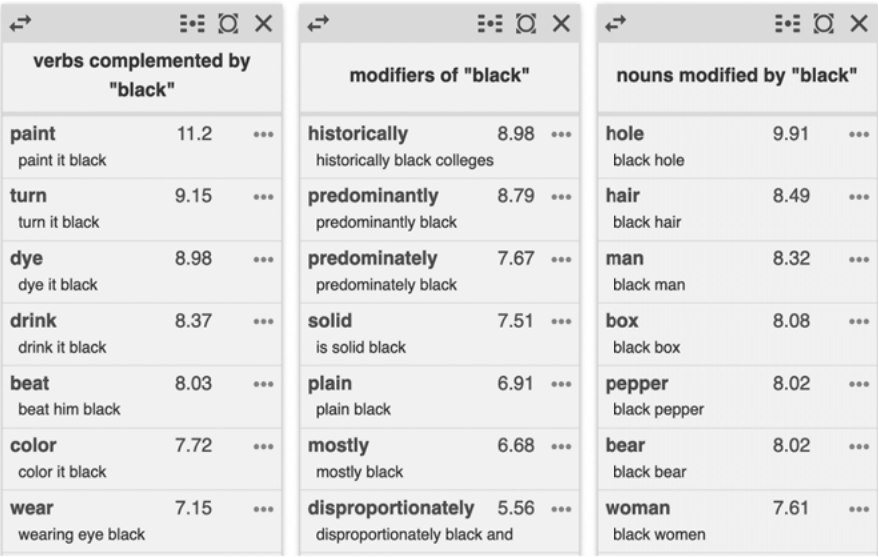


Abb. 2: Kollokationsanalyse von *schwarz* im DeReKo 2020-I

Die Rangfolge wird mit dem für die Ermittlung von Kollokationen einschlägigen statistischen Maß der log-likelihood-ratio (LLR) ermittelt. Kollokationsanalysen sind für lexikografische Zwecke sehr nützlich und können insbesondere wichtige Daten für die Phraseologieforschung liefern. Der idiomatische Ausdruck *schwarzes Schaf* und die Kollokationen *schwarze Zahl* und *schwarze Null* finden sich unter den Ergebnissen mit den höchsten log-likelihood Werten.

Mit Hilfe der Webanwendung *Sketch Engine* (Kilgariff et al. 2014) können Kollokationsanalysen sprachvergleichend durchgeführt werden. Sketch Engine hält eine große Anzahl von elektronischen Korpora für eine Vielzahl von Sprachen bereit, darunter auch sehr große Textsammlungen, die aus Webdaten bestehen.

Wenn man die Kollokate von *black* mit denen von *schwarz* im Deutschen vergleicht, fallen linguistische und kulturelle Gemeinsamkeiten und Unterschiede gleichermaßen auf. So tritt die Bedeutung von *black* als ethnische Bezeichnung viel stärker hervor als bei *schwarz* im Deutschen. Dies zeigt sich besonders bei den Kollokaten von Adverbien. Bei den Mehrwortlexemen fallen sowohl Gemeinsamkeiten wie *black hole* und *schwarzes Loch*, aber auch Unterschiede wie *schwarze Null* oder *schwarze Zahlen* auf, die keine direkten Entsprechungen im Englischen haben.



verbs complemented by "black"		
paint	paint it black	11.2
turn	turn it black	9.15
dye	dye it black	8.98
drink	drink it black	8.37
beat	beat him black	8.03
color	color it black	7.72
wear	wearing eye black	7.15

modifiers of "black"		
historically	historically black colleges	8.98
predominantly	predominantly black	8.79
predominately	predominately black	7.67
solid	is solid black	7.51
plain	plain black	6.91
mostly	mostly black	6.68
disproportionately	disproportionately black and	5.56

nouns modified by "black"		
hole	black hole	9.91
hair	black hair	8.49
man	black man	8.32
box	black box	8.08
pepper	black pepper	8.02
bear	black bear	8.02
woman	black women	7.61

Abb. 3: Datenauszug aus dem Word Sketch für das Adjektiv *black*

Für das Englische stellen die digitale Korpussammlung und die zugehörigen Such- und Visualisierungswerkzeuge, die von Mark Davies auf der Webseite English-Corpora.org angeboten werden, eine exzellente Ressource für linguistische Untersuchungen dar. Sie bietet eine Übersicht über neun Korpora unterschiedlicher Größe, geografischer und zeitlicher Abdeckung sowie unterschiedlicher Textsorten. Besonders erwähnenswert sind das *iWeb Corpus* (Davies/Kim 2019), das *Corpus of Contemporary American English* (COCA; Davies 2009) und das *Corpus of Historical American English* (COHA; Davies 2010). Diese Korpora können entweder direkt auf den eigenen Rechner heruntergeladen oder mit Hilfe des Suchinterfaces im Netz recherchiert werden. Das Suchinterface ist außerordentlich vielseitig und daher für ganz unterschiedliche linguistische Fragestellungen äußerst nützlich. Das weitaus größte Korpus, das von Mark Davies zur Verfügung gestellt wird, ist das iWeb Corpus, das englische Daten aus 6 verschiedenen Ländern umfasst.

Dank der *Corpora from the Web Initiative* (COW; Schäfer 2016) von Felix Bildhauer und Roland Schäfer gibt es Webkorpora für verschiedene europäische Sprachen, die für sprachvergleichende Studien verwendet werden können und mit Hilfe der SketchEngine durchsuchbar sind.

Mit dem deutschen Textarchiv (DTA; Geyken et al. 2018) steht ein hochwertiges Korpus mit diachronen Daten des Deutschen zur Verfügung. Das DTA ist ein Langzeitvorhaben an der Berlin-Brandenburgische Akademie der Wissenschaften. Es umfasst Texte des frühen 16. bis zum frühen 20. Jahrhundert aus vier verschiedenen Textsorten: literarische und wissenschaftliche Texte, Zeitungstexte und Gebrauchsliteratur. Mit der linguistischen Suchmaschine DDC lässt sich das DTA online durchsuchen.

Als Beispiel, wie sich das DTA für linguistische Untersuchungen nutzen lässt, möchte ich eine Studie zur morphologischen Produktivität von *-bar* Adjektiven (Hinrichs 2017) vorstellen, die ich auf Grundlage der DTA Daten durchgeführt habe.

Wenn man die Zeitspanne von 1600 bis Ende des 19. Jahrhunderts, die das DTA abdeckt, so zeigt sich, dass die Produktivität des *–bar* Suffixes in diesem Zeitraum signifikant zugenommen hat. Das kann man u.a. anhand der Verlaufskurve in Abbildung 4 ablesen, die die Vorkommenshäufigkeit pro 1 Million Wörter zeigt.

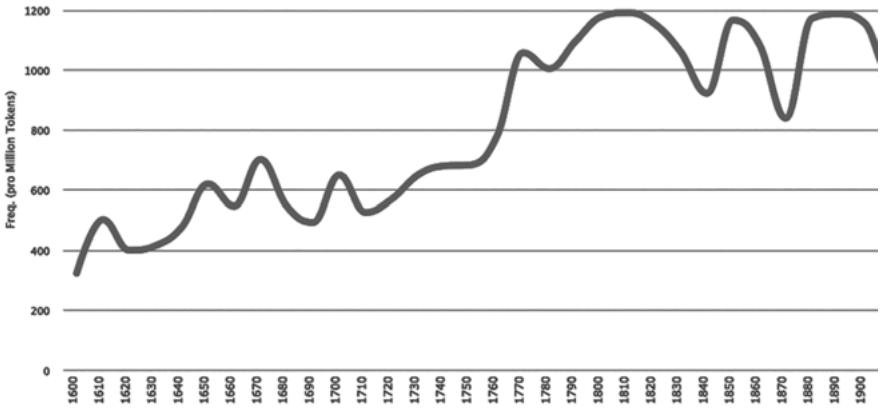


Abb. 4: Relative Häufigkeit von *-bar* Adjektiven pro 1 Million Token im DTA

Nun könnte man natürlich vermuten, dass die relative Häufigkeit von anderen Derivationssuffixen wie *-ig, -lich* und *-isch* für Adjektivbildungen im gleichen Zeitraum ebenfalls zunimmt. Dann hätten wir es mit einem allgemeinen Zuwachs an Produktivität zu tun. Dies ist aber nicht der Fall, wie die Abbildung 5 zeigt: Die relative Häufigkeit von Adjektiven, die auf *-lich* enden, bleibt im DTA für den Vergleichszeitraum relativ konstant. Dieser Befund gilt auch für Adjektivbildungen auf *-ig* und *-isch*. Der Zuwachs an Produktivität ist vielmehr ein Spezifikum der Adjektive mit dem *-bar* Suffix.

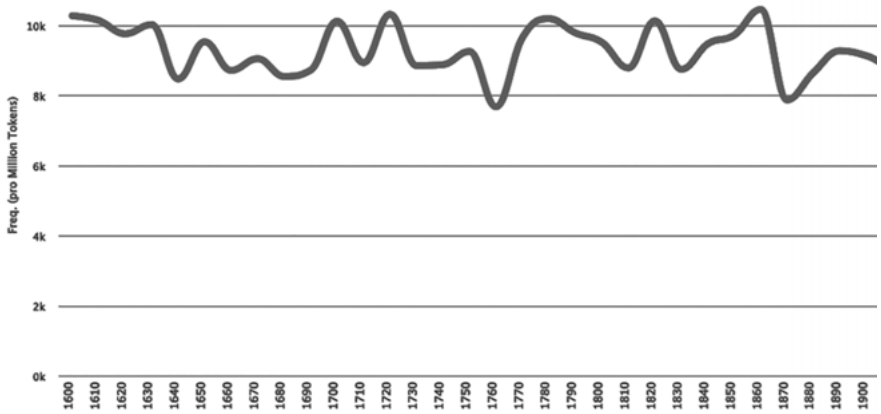


Abb. 5: Relative Häufigkeit von *-lich* Adjektiven pro 1 Million Token im DTA

Zusätzliche Evidenz für die Hypothese einer gesteigerten Produktivität des *-bar* Suffixes bietet das Vorkommen von Hapax Legomena im untersuchten Zeitraum. Die Signifikanz von Hapax Legomena für ein Ansteigen von morphologischer Produktivität geht auf die Studien von Baayen und Lieber (1991) zurück. Wenn man die drei Jahrhunderte der DTA Texte in Zeitscheiben von jeweils zwei Dekaden (siehe Abb. 6) unterteilt, so steigt die Anzahl der Hapax Legomena von etwa 10 bis auf fast dreihundert an Ende des 19. Jahrhunderts.

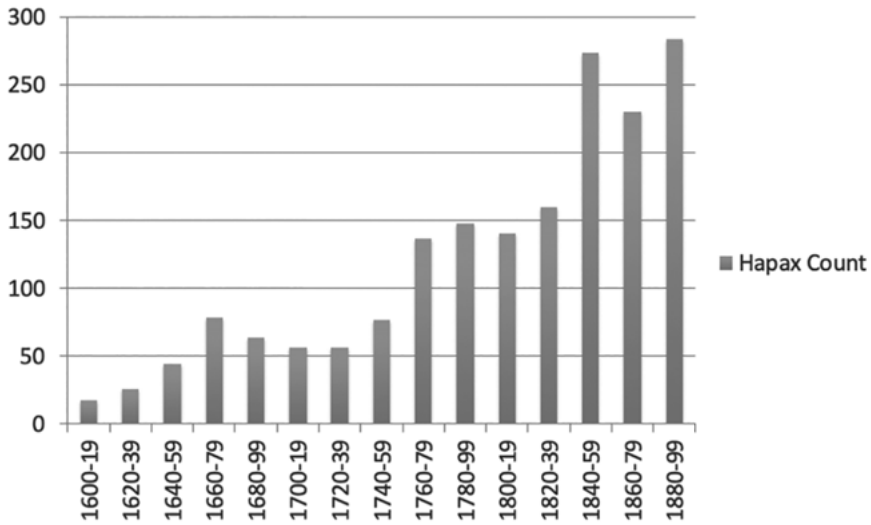


Abb. 6: Hapax Legomena für bar-Adjektive im DTA

Flury (1964) hat darauf hingewiesen, dass die Bedeutung des Suffix -bar auf MHD bære zurückgeht – verwandt mit dem Englischen bear (‚tragen'). D. h. ein ehrbarer Mensch bezeichnet eine Person, die Zeichen der Ehre trägt. Ein dankbarer Mensch war jemand, der Zeichen der Geneigtheit hervorbringt, der Dank bringt oder ausdrückt. Wenn man sich die häufigsten -bar Adjektive im DTA zu Beginn des 17. Jahrhunderts anschaut, so finden sich darunter ausschließlich denominale -bar Adjektive und kein einziges deverbales Adjektiv. Die Bedeutung von -bar hat sich dann zunehmend zu einer abstrakteren, oft dispositionalen Bedeutung gewandelt. Aufgrund der Formengleichheit von nominalen und verbalen Stämmen wie dank- ist dann die Möglichkeit der Re-Analyse und eine Generalisierung auf verbale Basiswörter und damit einhergehend eine ansteigende Produktivität gegeben.

3 Baumbanken für das Deutsche und für sprachvergleichende Untersuchungen

Linguistisch annotierte Textkorpora haben sich traditionell auf Annotationen auf der Wortebene beschränkt: auf die Lemmatisierung von Wortformen, auf die Annotation von Morphologie und Morphosyntax. Seit den 1980er Jahren sind zunehmend auch syntaktische Annotationen hinzugekommen und sogenannte *Baumbanken* entstanden. Für das Gegenwartsdeutsche stehen inzwischen vier Baumbanken zur Verfügung (vgl. Tab. 1), die auf schriftsprachlichen Daten beruhen und die manuell annotiert worden sind.

Tab. 1: Übersicht über Baumbanken für Texte der deutschen Gegenwartssprache

	Anzahl der annotierten Sätze	Art der Annotation	Textsorte
Hamburg Dependency Treebank	206.794 Sätze	Dependenz-Grammatik	IT Nachrichten on-line
Tübinger Baumbank TüBa-D/Z	104.787 Sätze	Phrasenstrukturgrammatik	Zeitungstexte
Tiger	50.000 Sätze	Phrasenstrukturgrammatik	Zeitungstexte
Negra	20.602 Sätze	Phrasenstrukturgrammatik	Zeitungstexte

Die Hamburg Dependency Treebank (Foth et al. 2014) umfasst über 200.000 annotierte Sätze eines online Services von IT-Nachrichten. Die Annotation folgt den syntaktischen Strukturen der Dependenzgrammatik. Die Tübinger Baumbank TüBa-D/Z ist mit circa 100.000 Sätzen die größte Baumbank des Deutschen mit phrasenstrukturgrammatischen Annotationen. Die TüBa-D/Z (Telljohann et al. 2004; Telljohann et al. 2017) und die TIGER Baumbank (Brants et al. 2004), eine weitere vielverwendete Baumbank des Deutschen, basieren jeweils auf Zeitungstexten: die TüBa-D/Z auf der taz und die TIGER Baumbank auf einem Frankfurter-Rundschau-Korpus.

Die Webanwendung TüNDRA (Tübingen annotated Data Retrieval Application; Martens 2013) stellt ein Such- und Visualisierungswerkzeug für gegenwärtig mehr als fünfzig Baumbanken für verschiedene Sprachen zur Verfügung. Abbildung 7 präsentiert eine TüNDRA Visualisierung für den in der TüBaD/Z Baumbank enthaltenen Satz *Veruntreute die AWO Spendengelder?*.

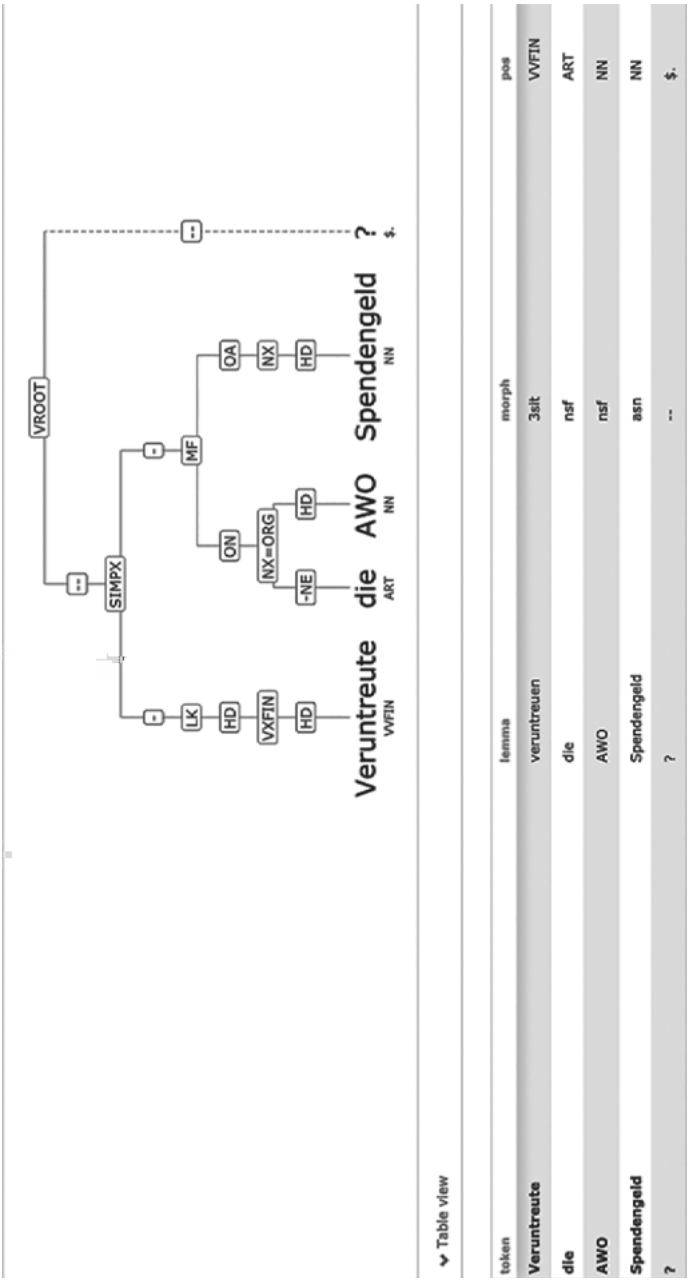


Abb. 7: TüNDRA Visualisierung der TüBa-D/Z Morphologie und Konstituentenstruktur

Die linguistischen Annotationen auf der Wortebene werden in Tabellenform dargeboten. Die Baumstruktur besteht aus Knoten, die die syntaktischen Kategorien anzeigen und topologische Felder wie LK (für: linke Klammer) und MF (für: Mittelfeld) identifizieren. Die rosa markierten Kanten wie z.B. HD zeigen die syntaktischen Köpfe einer Phrase an oder grammatische Funktionen wie ON (für: Subjekt) und OA (für: Akkusativobjekt).

In einer dependenzgrammatischen Struktur für den gleichen Satz (siehe Abb. 8) werden die Relationen zwischen Einzelwörtern durch Kantenbeschriftungen wie SUBJ, Akkusativobjekt und Determinierer identifiziert. Der Wurzelknoten ROOT zeigt auf den Kopf der gesamten Struktur, im vorliegenden Satz auf das finite Verb *veruntreute*.

Eine interessante Frage, die sich bei annotierten Korpora immer wieder stellt, betrifft die Anzahl und Auswahl von Daten, die für Untersuchungen sprachlicher Phänomene benötigt werden. Wenn man sich für hochfrequente Wörter oder syntaktische Konstruktionen interessiert, sind dafür vergleichsweise weniger Daten erforderlich, als wenn man sich für relativ seltene linguistische Phänomene interessiert. Ein relativ seltenes syntaktisches Phänomen ist die sogenannte *Zwischenstellung* von finiten Auxiliaren im Verbalkomplex im Deutschen. Ein einschlägiges Beispiel zeigt Abbildung 9 mit dem Satz *Unklar ist jedoch, ob Stephane Henchoz mitspielen wird können*, wo das finite Auxiliar *wird* zwischen dem Verb *mitspielen* und dem Auxiliar *können* steht.

Wenn man nach derartigen Beispielen in der manuell annotierten Baumbank TüBaD/Z sucht, wird man nicht fündig. Die TüBa-D/Z umfasst allerdings nur einen kleinen Datenausschnitt von etwas mehr als 100.000 Sätzen aus der taz Wissenschafts-CD mit 11,65 Millionen Sätzen. Das gesamte Korpus haben wir in Tübingen ebenfalls syntaktisch annotiert und stellen es unter dem Namen TüPP-D/Z zur Verfügung. Allerdings handelt es sich dabei um automatische Annotationen eines finite-state Parsers. Diese Annotationen haben daher nicht die gleiche Annotationsqualität wie manuell annotierte Daten. In der TüPP-D/Z (Müller 2004) finden sich 92 Belegsätze für die Zwischenstellung im deutschen Verbalkomplex, darunter auch der in Abbildung 10 gezeigte Satz.

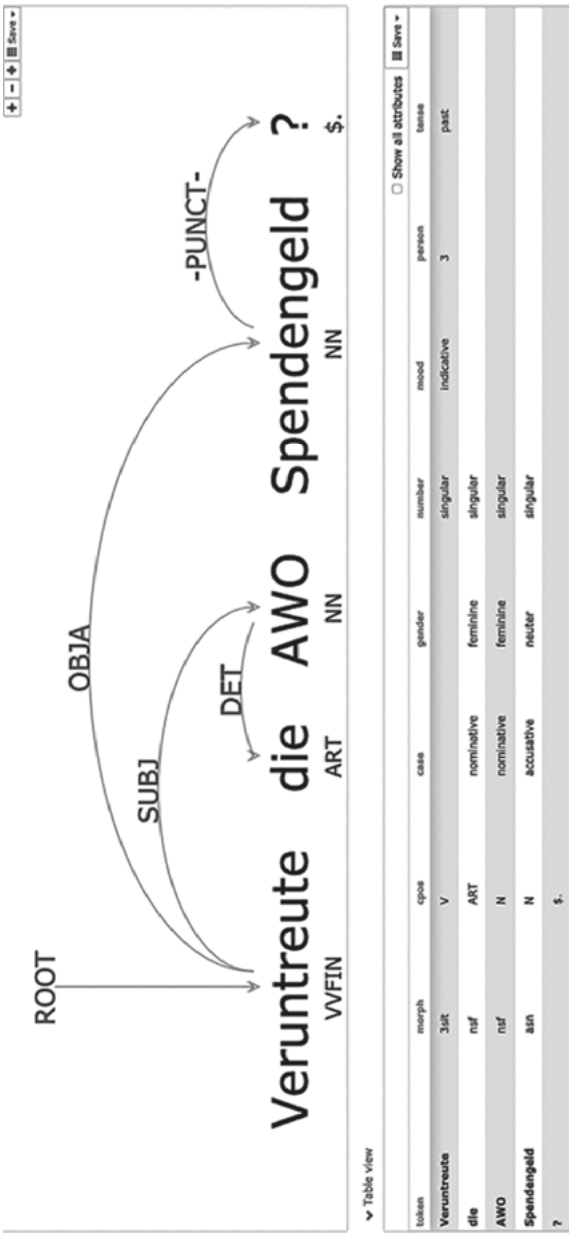


Abb. 8: TüNDRA Visualisierung der TüBa-D/Z Morphologie und Dependenzstruktur

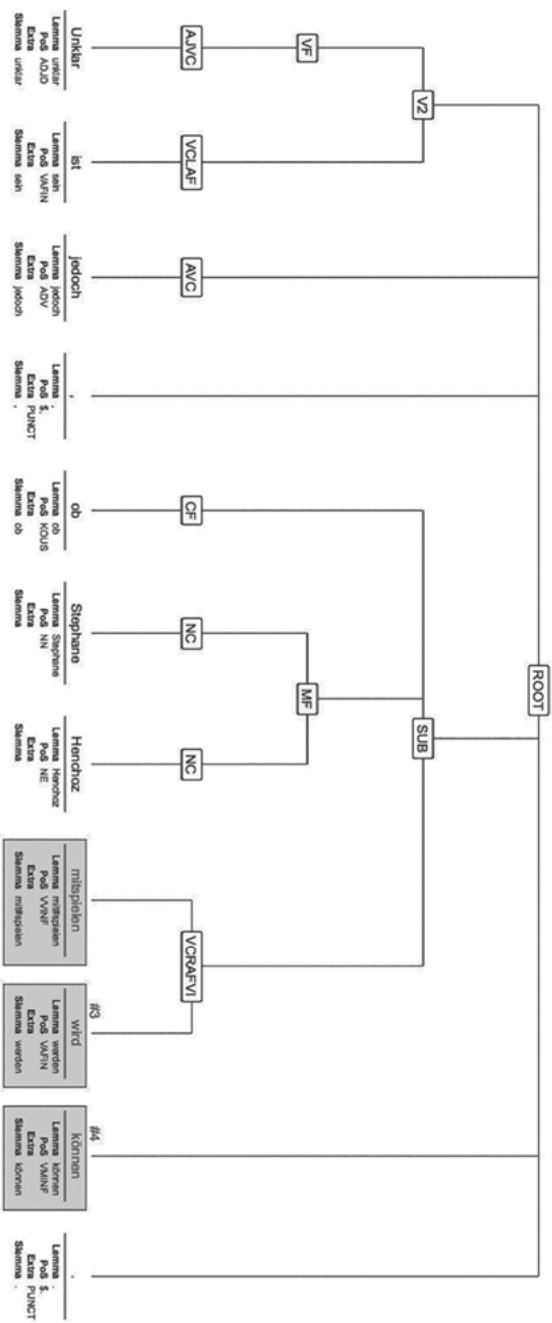


Abb. 9: Belegstelle für eine Zwischenstellung im Verbalkomplex aus der TüPP/D-Z Baumbank

Welche generelleren Einsichten lassen sich aus diesem Beispiel gewinnen? Zum einen die wichtige Warnung, dass man aus fehlenden Treffern in einem Korpus nicht ableiten darf, dass die gesuchte Konstruktion offenbar ungrammatisch ist. Zum anderen den Hinweis, dass man sich bei selteneren Phänomenen nicht auf kleine Datenmengen beschränken darf, sondern große Korpussammlungen benötigt, die nur mit automatischen computerlinguistischen Verfahren annotiert werden können und deren Annotationen weniger zuverlässig sind.

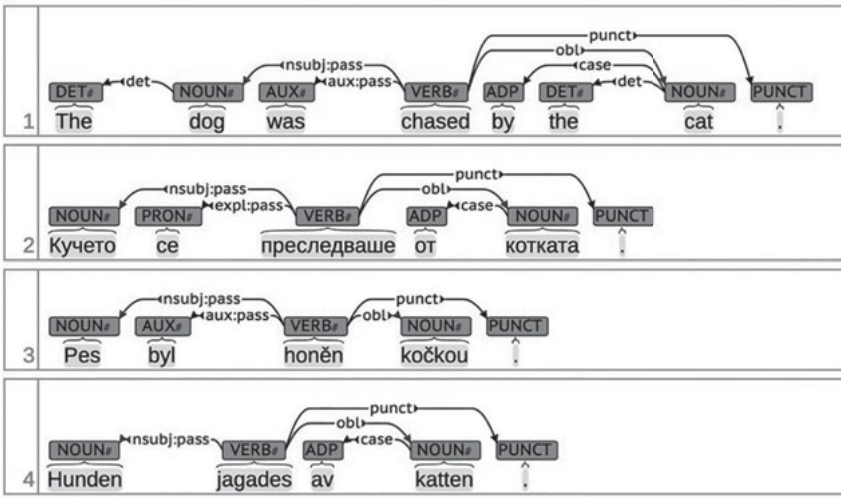


Abb. 10: Sprachübergreifende syntaktische Annotation nach den Richtlinien der Universal Dependencies; Quelle:universaldependencies.org/introduction.html (Stand: 8.11.2020)

Da es inzwischen Baumbanken für eine Vielzahl von Sprachen gibt, lassen sie sich auch für sprachvergleichende Untersuchungen nutzen. Allerdings werden solche sprachübergreifenden Studien oft durch unterschiedliche Annotationsschemata in den einzelnen Baumbanken erschwert. Um diese Schwierigkeit zu überwinden, hat sich vor circa zehn Jahren eine internationale Initiative mit dem Namen *Universal Dependency Treebanks* konstitutiert, deren Ziel es ist, Baumbanken für eine möglichst große Anzahl von Einzelsprachen nach einem einheitlichen Annotationsschema einer Universal Dependency Grammar (Nivre et al. 2016) zu annotieren. Abbildung 10 illustriert das Annotationsschema anhand eines Beispielsatzes aus dem Englischen, des Bulgarischen, des Tschechischen und des Schwedischen. Die Vereinheitlichung betrifft nicht nur die syntaktische, sondern auch die morpho-syntaktische Analyse, die nach einem Universal Tagset (Petrov et al. 2012) erfolgt.

4 Die CLARIN Forschungsdateninfrastruktur

Um Sprachressourcen nachhaltig für die Forschung bereitzustellen, bedarf es einer verlässlichen technischen Infrastruktur. Eine solche Infrastruktur hat sich auf nationaler und europäischer Ebene als ein Netzwerk von zertifizierten Datenzentren unter dem Namen CLARIN (Hinrichs/Krauwer 2014) etabliert. CLARIN ist das Akronym für *Common Language Resource and Technology Infrastructur*e. Auf europäischer Ebene hat sich CLARIN als Bestandteil der ESFRI Roadmap von europäischen Forschungsdateninfrastrukturen mit inzwischen über 20 Mitgliedsländern in Europa etabliert. Deutschland gehört zu den Gründungsmitgliedern des CLARIN. Das BMBF fördert mit dem CLARIN-D Projekt (Hinrichs/Trippel 2017) ein nationales Netzwerk von zertifizierten Datenzentren und Rechenzentren, die in das europäische CLARIN Netzwerk integriert sind und auf der Karte in Abbildung 11 angezeigt werden. Eines dieser Datenzentren ist am IDS in Mannheim angesiedelt und ein weiteres an der Universität Tübingen.

Abb. 11: Das Netzwerk von CLARIN Zentren in Deutschland

Die CLARIN Infrastruktur stellt Sprachressourcen der geschriebenen und gesprochenen Sprachen für zahlreiche Sprachen zur Verfügung, darunter auch Sprachen außerhalb Europas und bedrohte Sprachen. Darüber hinaus bietet CLARIN Werkzeuge für die automatische Annotation von Textkorpora an, darunter das in CLARIN-D entwickelte Werkzeug WebLicht (Hinrichs et al. 2010). WebLicht ist eine Webanwendung, mit der man seine eigenen Korpora für die Annotation von Texten hochladen kann und dann automatisch annotieren lassen kann. WebLicht (verfügbar unter: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page) bietet Annotationswerkzeuge für zahlreiche Sprachen an, darunter Deutsch, Englisch, Französisch, Italienisch, Niederländisch, Tschechisch und Ungarisch. Die in WebLicht verfügbaren Webdienste unterstützen mit der Tokenisierung von Wörtern und mit der Erkennung von Satzgrenzen die automatische Vorverarbeitung von Texten. Für die linguistische Analyse stehen Webservices für die Lemmatisierung, die Wortklassenerkennung, die morphologische Analyse, das syntaktische Parsing und die Erkennung von Eigennamen zur Verfügung.

Nutzende können ihre eigenen Texte in der WebLicht Webanwendung hochladen, um sie in WebLicht automatisch annotieren zu lassen. Zu diesem Zweck bietet WebLicht vordefinierte Verarbeitungsketten an, die von den Nutzenden per Knopfdruck ausgeführt werden können. Alternativ bietet WebLicht den Nutzenden die Möglichkeit, individuelle Webservices zu eigenen Verarbeitungsketten zusammenzustellen. Abbildung 12 zeigt die WebLicht Ausgabe für die automatische Erkennung von Eigennamen am Beispiel des Wikipedia Eintrags für den deutschen Physiker und Erfinder Hans Geiger. Die Annotationen, die mit Hilfe von WebLicht erzeugt werden, lassen sich mit der Such- und Visualisierungsplattform TüNDRA (verfügbar unter: https://weblicht.sfs.uni-tuebingen.de/Tundra/) visualisieren.

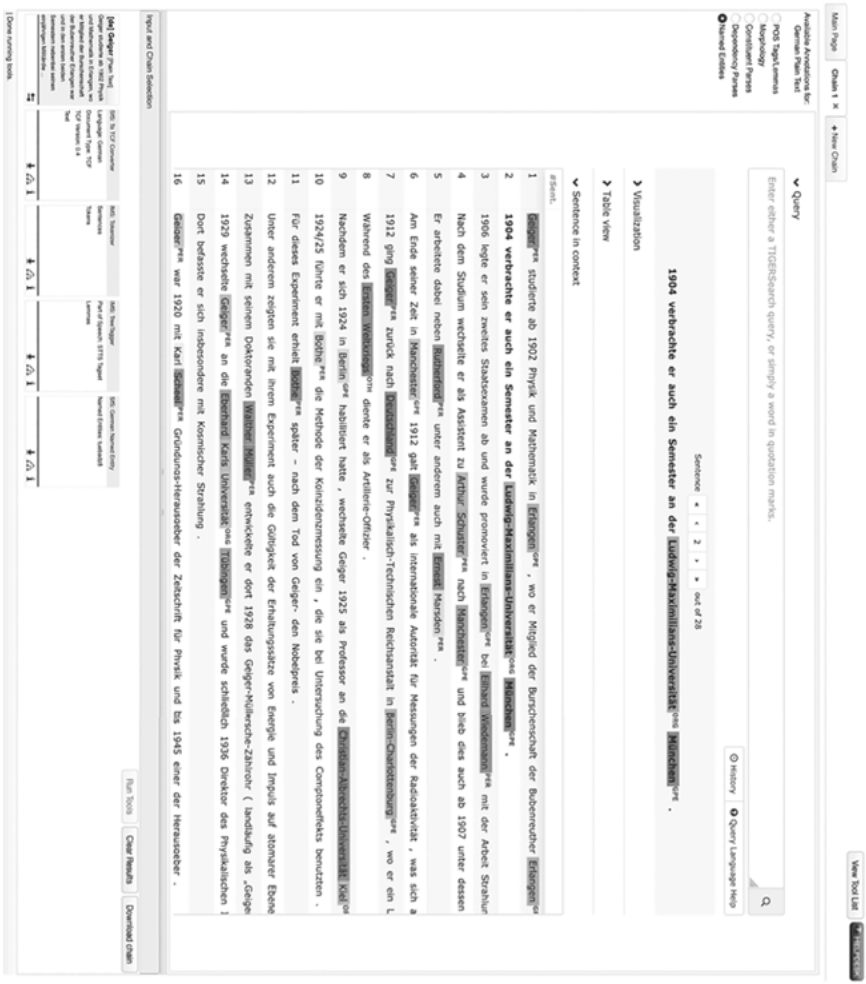


Abb. 12: Automatische Eigennamenerkennung mit WebLicht

5 Zusammenfassung und Ausblick

Viele der hier vorgestellten Sprachressourcen werden im Rahmen der CLARIN Infrastruktur nachhaltig zur Verfügung gestellt. Die Korpora sind entweder durch Suchportale recherchierbar oder werden per Download zur Verfügung gestellt, soweit dies mit den Copyrightbeschränkungen vereinbar ist. Die in CLARIN verfügbaren Annotationswerkzeuge wie WebLicht sowie die Such- und Visualisierungswerkzeuge werden, wann immer möglich, als Webanwendungen zur Verfügung gestellt, damit sie von den Nutzenden in einem Webbrowser verwendet werden können. Diese Philosophie einer technisch niederschwelligen Bereitstellung vermeidet die Notwendigkeit, Werkzeuge auf dem eigenen Rechner installieren zu müssen. Solche Installationen bringen häufig hohe technische Anforderungen mit sich in Form von Beschränkungen auf bestimmte Betriebssysteme oder auf bestimmte Versionen von Programmiersprachen. Dies stellt besonders für Nutzende aus den Geisteswissenschaften, die in vielen Fällen nicht über das dafür nötige technische Vorwissen verfügen, eine hohe Hürde für die Verwendung von digitalen Sprachressourcen dar.

In diesem Beitrag habe ich mich auf den Mehrwert von digitalen Textkorpora für die sprachwissenschaftliche Forschung konzentriert. Ich möchte abschließend auf drei weitere Gegenstandsbereiche verweisen, in denen die Linguistik gleichermaßen von digitalen Ressourcen profitieren kann. Für die Lexikografie bieten eine Vielzahl digitaler Wörterbücher eine hervorragende Datengrundlage. Für das Deutsche sind das Digitale Wörterbuch der deutschen Sprache (DWDS; Klein/Geyken 2010) und die umfangreichen Online-Ressourcen der Abteilung Lexik am IDS zu nennen. Neben digitalen Wörterbüchern stehen für sprachvergleichende Untersuchungen digitale Wortnetze zur Verfügung, die lexikalische und konzeptuelle Relationen zwischen Wortbedeutungen in einer Graphstruktur abbilden. Ein solches Wortnetz wurde mit dem Princeton WordNet (Fellbaum 1998) zuerst für das Englische entwickelt. Inzwischen sind zahlreiche Wortnetze für weitere Einzelsprachen entstanden, die durch einen interlingualen Index miteinander verknüpft sind. Dazu gehört auch das deutsche Wortnetz GermaNet (Feldweg/Hamp 1997; Henrich/Hinrichs 2010). Das Leipziger Wortschatzportal (Goldhahn/Eckart/Quasthoff 2012) stellt korpusbasierte Wörterbücher für gegenwärtig 252 verschiedene Einzelsprachen online zur Verfügung und bietet somit ebenfalls eine hervorragende Grundlage für sprachvergleichende Untersuchungen.

Die Dokumentation bedrohter Sprachen profitiert ebenfalls von der Bereitstellung digitaler Ressourcen. Für diesen Bereich möchte ich exemplarisch das DOBES Archiv (Wittenburg et al. 2002) nennen, das einen Online-Zugang zu digitalen Materialien der gesprochenen und geschriebenen Sprache für mehr als 60 bedrohte Sprachen ermöglicht. Für die Spracherwerbsforschung bietet das *Child*

Language Data Exchange System (CHILDES; MacWhinney 1991) eine digitale Ressource an, die bereits seit den 1980er Jahren kontinuierlich gewachsen ist und bis dato von über 4500 Studien weltweit zitiert worden ist. CHILDES gehört inzwischen zu den wichtigsten Referenzquellen in der Psycholinguistik.

Literatur

Baayen, Harald/Lieber, Rochelle (1991): Productivity and English Derivation. A Corpus Based Study. Linguistics 29, 5, S. 801–843.

Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. In: SPRACHREPORT 3/2005, S. 2–5.

Brants, Sabine/Dipper, Stefanie/Eisenberg, Peter/Hansen-Schirra, Silvia/König, Esther/Lezius, Wolfgang/Rohrer, Christian/Smith, George/Uszkoreit, Hans (2004): TIGER: Linguistic interpretation of a German Corpus. In: Journal of Language and Computation, 2, 4, S. 597–620.

Davies, Mark (2009): The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture, and linguistic insights. In: International Journal of Corpus Linguistics 14, 2, S. 159–190.

Davies, Mark (2010): The 400 million word corpus of Historical American English (1810–2009). In: Hegedus, Irén/Fodor, Alexandra (Hg.): English Historical Linguistics 2010. Selected papers from the sixteenth international conference on English Historical Linguistics (ICEHL 16). (= Current Issues in Linguistic Theory 325). Amsterdam/Philadelphia: Benjamins, S. 231–262.

Davies, Mark/Kim, Jong-Bok (2019): The advantages and challenges of "big data". Insights from the 14 billion word iWeb corpus. In: Linguistic Research 36, 1, S. 1–34.

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: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož. Paris: ELRA, S. 3586–3591.

Fellbaum, Christiane (1998): WordNet: An electronic lexical database. Cambrigde: MIT Press.

Flury, Robert (1964): Struktur- und Bedeutungsgeschichte des Adjektiv-Suffixes *-bar*. Univ. Diss. Zürich. Winterthur: Keller.

Foth, Kilian A./Köhn, Arne/Beuck, Niels/Menzel, Wolfgang (2014): Because size does matter: The Hamburg dependency treebank. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Loftsson, Hrafn/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik. Paris: ELRA, S. 2326–2333.

Geyken, Alexander/Boenig, Matthias/Haaf, Susanne/Jurish, Bryan/Thomas, Christian/Wiegand, Frank (2018): Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hg.): Digitale Infrastrukturen für die germanistische Forschung (= Germanistische Sprachwissenschaft um 2020 6). Berlin/Boston: De Gruyter, S. 219–248.

Goldhahn, Dirk/Eckart, Thomas/Quasthoff, Uwe (2012): Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Dogan, Mehmet Ugur/Maegaard, Bente/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul. Paris: ELRA, S. 759–765.

Hamp, Birgit/Feldweg, Helmut (1997): GermaNet – a lexical-semantic net for German. In: Proceedings of the ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid, 1997.

Henrich, Verena/Hinrichs, Erhard (2010): GernEdiT – The GermaNet editing tool. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Odijk, Jan/Piperidis, Stelios/Mariani, Joseph/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta. Paris: ELRA, S. 2228–2235.

Hinrichs, Erhard (2017): Morphological productivity of adjective formation in German – A diachronic corpus study using the CLARIN-D infrastructure. In: Book of abstracts CLARIN annual conference 2017, Budapest. Internet: https://www.clarin.eu/sites/default/files/Hinrichs-CLARIN2017_paper_19.pdf (Stand: 23.10.2020).

Hinrichs, Erhard/Krauwer Steven (2014): The CLARIN research infrastructure: Resources and tools for E-humanities scholars. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), S. 1525–1531. Internet: http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf (Stand: 28.10.2020).

Hinrichs, Erhard/Trippel, Thorsten (2017): CLARIN-D: Eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. In: Bibliothek – Forschung und Praxis 41, 2, S. 45–54.

Hinrichs, Erhard/de Kok, Daniël/Çöltekin, Çağrı (2015): Treebank data and query tools for rare syntactic constructions. In: Dickinson, Markus/Hinrichs, Erhard/Patejuk, Agnieszka/Przepiórkowski, Adam: Proceedings of the 13th International Workshop on Treebanks and Linguistic Theories (TLT14), Warschau, Dezember 2015. Warschau: Polish Academy of Sciences, S. 106–118.

Hinrichs, Erhard/Hinrichs, Marie/Zastrow Thomas (2010): WebLicht: Web-Based LRT Services for German. In: Kübler, Sandra (Hg.): Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010), Uppsala. Stroudsburg: ACL, S. 25–29. Internet: https://www.aclweb.org/anthology/P10-4005.pdf (Stand:26.10.2020).

Kilgarriff, Adam/Kilgarriff, Adam/Baisa, Vít/Bušta, Jan/Jakubíček, Miloš/Kovář, Vojtěch/Michelfeit, Jan/Rychlý, Pavel/Suchomel, Vít (2014): The sketch engine: ten years on. In: Lexicography 1, 1, S. 7–36.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich/Gouws, Rufus Hjalmar/Herbst, Thomas/Lobenstein-Reichmann, Anja/Schierholz, Stefan J./Schweickard, Wolfgang (Hg.): Lexikographica. International Annual for Lexicography, 26. Berlin/New York: De Gruyter, S. 79–93.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German reference corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta. Paris: ELRA, S. 1848–1854. Internet: http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (Stand:26.10.2020).

MacWhinney, Brian (1991): The CHILDES project. Tools for analyzing talk. Hillsdale, NJ.: Erlbaum.

Martens, Scott (2013): TüNDRA: A Web Application for Treebank Search and Visualization. In: Kübler, Sandra/Osenova, Petya/Volk, Martin (Hg.): Proceedings of the 12th Workshop on Treebanks and Linguistic Theories (TLT 12). Sofia: Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, S. 133–144.

Müller, Frank Henrik (2004): Stylebook for the Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z). SfS Technical Report: Seminar für Sprachwissenschaft. Tübingen: Universität Tübingen. Internet: https://sfs.uni-tuebingen.de/tupp/doc/stylebook.pdf (Stand: 26.10.2020).

Nivre, Joakim/de Marneffe, Marie-Catherine/Ginter, Filip/Goldberg, Yoav/Hajič, Jan/Manning, Christopher D./McDonald, Ryan/Petrov, Slav/Pyysalo, Sampo/Silveira, Natalia/Tsarfaty, Reut/Zeman, Daniel (2016): Universal dependencies v1: A multilingual treebank collection. In: Calzolari, Nicoletta/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. 1659–1666.

Petrov, Slav/Das, Dipanjan/McDonald, Ryan (2012): A universal part-of-speech tagset. 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. 2089–2096. Internet: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf (Stand: 26.10.2020).

Schäfer, Roland (2016): CommonCOW: Massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In: Calzolari/Choukri/Declerck/Goggi/Grobelnik/Maegaard/Mariani/Mazo/Moreno/Odijk/Piperidis (Hg.): Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), S. 4500–4504.

Telljohann, Heike/Hinrichs, Erhard/Kübler, Sandra (2004): The TüBa-D/Z treebank. Annotating German with a context-free backbone. In: Lino, Maria Teresa/Xavier, Maria Francisca/Ferreira, Fatima/Costa, Rute/Silva,Raquel (Hg.): Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lissabon. Paris: ELRA, S. 2229–2232. Internet: http://www.lrec-conf.org/proceedings/lrec2004/pdf/135.pdf (Stand: 26.10.2020).

Telljohann, Heike/Hinrichs, Erhard/Kübler, Sandra/Zinsmeister, Heike/Beck, Kathrin (2017): Stylebook for the Tübingen Treebank of written German (TüBa-D/Z). SfS Technical Report: Seminar für Sprachwissenschaft. Tübingen: Universität Tübingen. Internet: www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1707.pdf (Stand: 26.10.2020).

Wittenburg, Peter/Mosel, Ulrike/Dwyer, Arienne (2002): Methods of language documentation in the DOBES project. In: González Rodríguez, Manuel/Suarez Araujo, Carmen Paz (Hg.): Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas. Paris: ELRA, S. 36–42.