Was ist und was kann ein Referenzkorpus?

Norbert Richard WOLF

Abstract

What is a reference corpus, and what can it do?

The 'Deutsche Referenzkorpus (DeReKo)' of the Mannheimer Institut für Deutsche Sprache currently contains over 28 billion words, and it is constantly being expanded. The sheer size of the corpus makes it impractical for researchers to analyze its entire content. On the other hand, the DeReKo offers the possibility of taking seriously the principle that every research project needs its own corpus – by acting as a 'reference corpus' that can be used in combination with special corpora. This paper addresses the question of whether a corpus should contain complete texts or only statistically relevant extracts; it also discusses the uses and necessity of 'small corpora'.

Keywords: corpus linguistics, reference corpus, small corpora, compensatory competence, communication of meaning

Der Internetauftritt des Mannheimer Instituts für deutsche Sprache (IDS) beantwortet die sicherlich drängende Frage „Was ist ‚Korpuslinguistik'" kurz und bündig (URL 1):

> „Das wissenschaftliche Programm der Korpuslinguistik ist es, geleitet durch die explorative Analyse von sehr großen Sammlungen natürlichsprachlicher Daten neue Einsichten in die Strukturen, Gesetzmäßigkeiten, Eigenschaften und Funktionen von Sprache zu erlangen."

Diese Definition enthält das Postulat, dass nur „sehr große[] Sammlungen" als Korpora gelten können. Das Institut für deutsche Sprache (IDS) verweist auf einem ‚Flyer' (URL 2) darauf, dass es auch Aufgabe des IDS ist, für die „empirische[] Grundlage für die germanistisch-sprachwissenschaftliche Forschung" zu sorgen.

> „Zu diesem Zweck unterhält das Institut seit 1964 eine umfangreiche elektronische Stichprobe deutschsprachiger Texte aus Gegenwart und jüngerer Vergangenheit: das so genannte Deutsche Referenzkorpus (DeReKo)."

Dieses DeReKo ist „mit über 28 Milliarden Wörtern die weltweit größte Sammlung elektronischer Korpora mit deutschsprachigen Texten aus Gegenwart und neuerer Vergangenheit“ (URL 2). Die IDS-Homepage spricht schon von „32,83 Milliarden Wörtern (Stand 01.10.2017)“ (URL 3). Wenn ich mehr als 28 Milliarden Wörter als Grundlage von Korpusrecherchen nehmen will, dann frage ich mich, wie weit es überhaupt möglich ist, so viele Wörter bzw. Wortformen zu durchsuchen; ich werde auf dieses Problem noch zurückkommen. Die IDS-Spezialisten scheinen sich des Problems der Korpusgröße bewusst zu sein, denn sie schreiben auch:

> „Die Korpora geschriebener Gegenwartssprache des IDS werden im Hinblick auf Umfang, Variabilität, Qualität und Aktualität akquiriert und erlauben in der Nutzungsphase über COSMAS II die Komposition virtueller Korpora, die repräsentativ oder auf spezielle Aufgabenstellungen zugeschnitten sind.“ (URL 1)

Das will wohl sagen, dass sich das DeReKo nicht als ein Korpus versteht, sondern als eine Textsammlung, aus der sich jede Person ein Korpus nach eigenen Bedürfnissen zusammenstellen kann. Streng genommen, gilt für jedes sprachwissenschaftliche Projekt, das korpusgestützt durchgeführt werden soll, das Postulat, dass für jedes einzelne Projekt ein eigenes Korpus erstellt werden muss. Das heißt natürlich auch, dass es ausreichen kann, zu prüfen, ob ein vorhandenes Korpus auch für das spezielle Projekt geeignet ist.

Dementsprechend informiert der IDS-Flyer auch darüber, dass das DeReKo „belletristische, wissenschaftliche und populärwissenschaftliche Texte, eine große Zahl von Zeitungstexten sowie eine breite Palette weiterer Textarten“ (URL 2) enthält.

> „Der große Vorteil von Korpora ist, dass sie nicht nur authentisches Sprachmaterial beinhalten, sondern darüber hinaus auch Informationen zur Häufigkeit und zur Verwendung von Wörtern, grammatischen Kategorien und anderen sprachlichen Einheiten liefern.“ (Scherer 2006:10)

Diese Äußerung Carmen Scherers macht auch deutlich, dass Korpora die Datengrundlage für die unterschiedlichsten sprachwissenschaftlichen Untersuchungen liefern können. Und – ich wiederhole variierend – die unterschiedlichen Untersuchungen erfordern unterschiedliche Korpora. Dies sei mit einem Beispiel illustriert: In den Jahren 1984 bis 1992 gab es an den Universitäten Würzburg und Eichstätt den Sonderforschungsbereich 226 ‚Wissensorganisierende und wissensvermittelnde Literatur im Mittelalter‘; dort leitete ich das Teilprojekt ‚Linguistische Probleme volkssprachlicher Wissensvermittlung‘, in dem wir die Wort- und Begriffsbildung in wissensliterarischen Texten beschreiben wollten (und auch beschrieben haben, vgl. Brendel/Frisch/Moser/Wolf 1997). In zwei weiteren Projekten, und zwar in Erlangen und in Bonn, wurde ebenfalls die Wortbildung im Frühneuhochdeutschen untersucht. Zu der Zeit, als wir mit unserer Arbeit begannen, gab es das sog. ‚Bonner Frühneuhochdeutschkorpus‘, das in den Jahren 1972 bis 1974 in erster Linie für die Fertigstellung der ‚Grammatik des Frühneuhochdeutschen‘ erstellt wurde. Die „Zusammenstellung erfolgte nach einem Bündel von Suchkriterien mit dem Ziel, eine räumliche, zeitliche und textartenbezogene Systematik der Quellengrundlage zu gewährleisten“ (Dammers/Hoffmann/Solms 1988:44). Für die ‚Grammatik des Frühneuhochdeutschen‘ wurden aus jedem Text „ca. 30 Normalseiten“ zu je „ca. 400 Wörtern“ (ebd. 50) ausgewählt. Dieses Korpus wurde dann auch dem Bonner Projekt ‚Verbableitung im Frühneuhochdeutschen‘ (vgl. Prell/Schebben-Schmidt 1996:9–12) zugrunde gelegt. Die Bonner Kolleginnen und Kollegen verstanden damals ihr Korpus als ein ‚Referenzkorpus‘, also als ein Korpus, „das dazu bestimmt ist, eine Sprache in ihrer Gesamtheit zu repräsentieren, und eine Vielzahl von sprachlichen Informationen zu liefern“ (Scherer 2006:27).

Bei unserem Würzburger Projekt mussten wir ganz anders verfahren. Wir brauchten kein Referenzkorpus, sondern ein ‚Spezialkorpus', das „eine bestimmte Varietät" (Scherer 2006:28) des Frühneuhochdeutschen, in unserem Fall die Sprache in ‚wissensliterarischen' Texten, wie sie Gegenstand des damaligen SFB waren, repräsentierte. Den untersuchten „Teilbereich der Sprache sollten" die „Spezialkorpora […] hinreichend repräsentieren" (ebd.). Wir wählten Texte der Wissensbereiche „‚Naturwissen', ‚Recht' und ‚theologisches Wissen'" sowie einen „überwiegend narrative[n] Text als Kontrollinstanz" (Grimm 1989:67). Alle Korpustexte waren Übersetzungen aus dem Lateinischen, als spezielles Kriterium kam noch der „Übersetzungsansatz" (ebd.) wörtliche oder sinngemäße Übersetzung hinzu.

Für uns war indes noch etwas ganz Anderes von Bedeutung: Wir wollten nicht kurze Auszüge zahlreicher Texte durchforsten, sondern umfangreiche Ganztexte. Grundlage für diese Entscheidung war eine Analyse der Verteilung von Wortbildungskonstruktionen im Text und über den Text. Ich bringe hier zwei Beispiele aus der ‚Summa Legum' des sog. Doctor Raymundus von Wiener-Neustadt, die um 1345 entstanden ist. Die beiden Tabellen sind gleich strukturiert: Die oberste waagrechte Reihe gibt zunächst die Zahl der gefundenen Suffixbildungen an und errechnet mit Hilfe der Seitenzahl der gedruckten Edition die durchschnittliche Frequenz, d.h. im ersten Fall begegnen sich (rein rechnerisch) auf jeder Seite zwei Bildungen mit dem jeweiligen Suffix. Der Text wird dann in Teile zu je 50 Seiten ‚zerlegt' (die ersten drei Spalten), die vierte Spalte gibt an, wie viele Belege (‚Tokens') sich tatsächlich finden; in der fünften Spalte finden wir die Angabe, wie viele Tokens rein rechnerisch zu erwarten wären, wenn die Belege gleichmäßig über den Text verteilt wären. Die Differenz zwischen dem tatsächlichen Vorkommen und der statistisch erwarteten folgt in der letzten Spalte.

Das erste Beispiel betrifft die Substantive, die mit dem Suffix *-ung* abgeleitet worden sind (nach Grimm 1989:70):

***-ung*-Substantive in der ‚Summa Legum'**					
1145 Belege	S. 123-686				2 Belege/Seite
von	**bis**	**Seitenzahl**	**Belege**	**erwartet**	**Differenz**
123	172	50	64	102	-38
173	222	50	71	102	-31
223	272	50	72	102	-30
273	322	50	102	102	0
323	372	50	68	102	-35
373	422	50	124	102	+22
423	472	50	111	102	+9
473	522	50	169	102	+67
523	572	50	122	102	+20
573	622	50	95	102	-7
623	672	50	94	102	-8

Tab. 1: Verteilung der *-ung*-Ableitungen über den ganzen Text

Allein auf den ersten 50 Seiten enthält der Text, statistisch formuliert, 38 Bildungen zu wenig, während „im letzten Dritten der ‚Summa Legum' Spitzen bis zu 169 Belege auf 50 Seiten" (Grimm 1989:70) vorkommen.

Als zweites Beispiel bringe ich die *-schaft*-Derivate in der ‚Summa Legum' (Grimm 1989:71):

-*schaft*-**Substantive in der ‚Summa Legum‘**					
191 Belege	S. 123-686 = 564 Seiten		0 Belege/Seite		
von	**bis**	**Seitenzahl**	**Belege**	**erwartet**	**Differenz**
123	172	50	9	17	- 8
173	222	50	21	17	+ 4
223	272	50	8	17	- 9
273	322	50	19	17	+ 2
323	372	50	56	17	+39
373	422	50	33	17	+16
423	472	50	9	17	- 8
473	522	50	6	17	-11
423	572	50	18	17	+ 1
573	622	50	11	17	- 6
623	672	50	0	17	-17

Tab. 2: Verteilung der -*schaft*-Ableitungen über den ganzen Text

Die ungleichmäßige Verteilung ist auch in diesem Fall durch die Gestaltung des Textinhaltes bedingt. Zugleich ist zu beachten, dass auf diese Weise nur ‚Belege‘, in der Sprache der Statistik ‚Tokens‘, erfasst werden und nicht ‚Typen‘ resp. ‚Types‘. Als Reaktion auf die Vorstellung des Würzburger Ansatzes auf einem Kolloquium im Oktober 1988 notiert Heinz-Peter Prell in einer Anmerkung (1989:44, Anm. 2):

> „In der Abschlußdiskussion des Kolloquiums wurde die Frage aufgeworfen, ob es nicht grundsätzlich vorzuziehen sei, auch umfangreiche Texte jeweils vollständig auszuwerten; so zeigen von den Würzburger Mitarbeitern erstellte Statistiken für bestimmte Phänomene sehr starke Schwankungen der Belegverteilung in verschiedenen Partien eines Textes. Derartige Untersuchungen werden sicher bei der Beurteilung von Belegfrequenzen zu berücksichtigen sein. Unser Interesse gilt jedoch zunächst der Erfassung von semantischen und morphologischen Typen der Wortbildung des Verbs sowie ihrer diachronen Entwicklung, in zweiter Linie den zu den einzelnen Typen belegten Lexemen und erst zuletzt der Frage der Beleghäufigkeit.“

Prell übersieht dabei, dass es durchaus sein kann, dass auch ‚Typen‘ unterschiedlich verteilt sind. Außerdem ist auch für eine ‚allgemeine‘ Wortbildungslehre von Belang, welche Typen funktional stark belastet sind und welche nur eine periphere Rolle spielen. Für die Beschreibung des wissensliterarischen Funktiolekts ist dies hingegen eine grundlegende Frage. Es spricht demnach Einiges dafür, dass Korpora für eine Reihe von Fragestellungen aus Ganztexten und nicht bloß aus einer Auswahl von Seiten bestehen sollten.

Zu einer Wortbildungsanalyse gehört auch die Frage der Motiviertheit der Bildungen. Zur Lösung dieses Problems können wir bei gegenwartssprachlichen Daten auf unsere muttersprachliche Kompetenz zurückgreifen. Doch bei historischen Sprachstufen wie dem Frühneuhochdeutschen müssen wir uns um eine „Ersatzkompetenz“ (Korhonen 1978:7) bemühen. Da wir auch seinerzeit die Motiviertheit von Wortbildungen mit der ‚Paraphrasemethode‘ nachzuweisen versuchten, wurde und wird es notwendig, auf die Kontexte der jeweiligen Wortbildung zurückzugreifen, denn ein „sicherer Nachweis einer Motivationsbeziehung kann nur dann erbracht werden, wenn die Basis“ im selben Text oder in anderen Texten desselben Autors oder in zeitgenössischen Texten „belegt ist“ (Habermann/Müller 1989:55). Der prototypische Nachweis einer Motivationsbeziehung zwischen Wortbildungsbasis und Wortbildungsprodukt ist der Beleg in ein und demselben Text, der häufig

genug auch Wortbildungsparaphrasen liefert. Für all das bedarf es umfangreich(er)er Ganztexte und nicht bloß einiger und zufällig ausgewählter Normalseiten.

Damit sind wir bei einer weiteren fundamentalen Frage angekommen, nämlich bei der, ob Korpuslinguistik immer eine „Analyse von sehr großen Sammlungen natürlichsprachlicher Daten" (URL 1) sein muss. Ich möchte eine Antwort auf diese Frage ebenfalls mit Hilfe eines früheren Projekts geben.

In den 1990er Jahren hatte der Würzburger Lehrstuhl für deutsche Sprachwissenschaft mit dem germanistischen Institut der Universität Jyväskylä ein Kooperationsprojekt zur kontrastiven Sprachwissenschaft Deutsch – Finnisch. Zu diesem Zweck wurde das kontrastive FinDe-Korpus erstellt, das aus zwei Modulen besteht, und zwar aus einem Übersetzungsmodul und einem Parallelmodul (ich weiß, dass ich mich von der üblichen Terminologie, die ein Übersetzungskorpus ‚Parallelkorpus' nennt, entferne). Das Übersetzungsmodul besteht aus drei original-deutschen Texten und ihren Übersetzungen ins Finnische sowie drei original-finnischen Texten und den Translaten ins Deutsche. Bei der Zusammenstellung war uns wichtig, immer Texte von verschiedenen Autoren und verschiedenen Übersetzern zu übernehmen. Das Parallelmodul bestand aus deutschen und finnischen Zeitungstexten zum selben Thema bzw. Ereignis, etwa zum Irakkrieg 1991. Wir gingen davon aus, dass die wesentlichen Quellen für die Zeitungsartikel englischsprachige Texte von den großen internationalen Nachrichtenagenturen waren. Damals gab es von einigen Zeitungen maschinenlesbare Ausgaben bzw. Jahrgangs-CDs.

Ein Teilprojekt des ganzen Unternehmens war die Beschreibung der Verba dicendi in den beiden Sprachen. Dabei begegneten wir Wortverwendungen wie dieser:

> *Eine kroatische Menschenrechtsgruppe warf der kroatischen Armeeführung derweil vor, Männer bosnischer Herkunft zum Kampfeinsatz in Bosnien zu zwingen. Ein Sprecher des Kroatischen Helsinki Komitees* ***erklärte****, seiner Organisation lägen Briefe von Familien der Männer vor, in denen beschrieben sei, wie die in Kroatien lebenden Männer aus Bosnien mit angedrohten Diziplinarmaßnahmen gegen ihren Willen zum Kampf in einer Einheit von „Freiwilligen" in den Reihen der bosnischen Kroaten gebracht würden. Das Verteidigungsministerium in Zagreb wies die Vorwürfe zurück*. (SZ 03.01.1994)

Wir wollten nun der Bedeutung von *erklären* in diesem Beispiel und anderen nachgehen. Vergleichbares war in finnischen Zeitungen nicht zu finden. Das Erste, was wir machten, waren Blicke in die einsprachigen Wörterbücher des Gegenwartsdeutschen, das Duden-Universalwörterbuch (Duden 2011) und das Wahrig-Wörterbuch der deutschen Sprache (Wahrig 2012). Keines dieser Wörterbücher bucht den Sprachgebrauch, wie wir ihn im Zeitungsartikel vom Januar 1994 finden, nämlich das Verb *erklären* mit direkter oder indirekter Rede, also mit einem Inhaltssatz als Objekt. Um die Bedeutung genauer beschreiben und die Wortverwendung fundiert erklären zu können, musste ich auf Korpora zurückgreifen. Ich wiederhole hier die Ergebnisse meiner Recherchen im Jahre 2009 (nach Wolf 2011):

Ich habe zunächst im damaligen DeReKo des Instituts für deutsche Sprache recherchiert und dort im Belletristik-Korpus, das aus 57 Texten bestand, die Form *erklärt** 673mal gefunden. Im Korpus ‚Berliner Morgenpost 1997-1999' kam *erklärt** 11.460mal vor. Diese Belegzahl ist ohne Zweifel eindrucksvoll; es dürfte aber jeden sinnvollen Aufwand übersteigen, etwas mehr als 12.000 Belege durchzusehen. Um eine etwas bessere Übersicht zu bekommen, habe ich auf das alte und bewährte Würzburger dtv-Korpus zurückgegriffen, das aus zehn dtv-Bänden, die im Jahre 1993 erschienen sind, besteht. Und daraus habe ich den Roman ‚Scheintod' von Eva Demski (1993) und das mythologische Lexikon ‚Who's who in der antiken Mythologie' von Gerhard Fink (1993) ausgewählt. Hier der Befund:

Das Verbum *erklären* kommt in diesen beiden Bänden insgesamt 36mal vor. Folgende Bedeutungen sind vertreten:

Bedeutung	**Frequenz im Korpus**	**im Roman**	**im Lexikon**
erklären			
‚erläutern'	21	15	6
‚äußern'	9	1	8
‚etw. als etw. bezeichnen	1	1	
Wendung *den Krieg erklären*	1		1
sich erklären			
‚Begründung finden'	3		3
‚einen Heiratsantrag machen'	1	2	

Tab. 3: Lesarten von *erklären* im Korpus

Dazu einige Beispiele:

‚erläutern'

> *Es hatte was Täppisches, immer, wenn irgend etwas passiert war, hinterherzudenken und einem Gericht zu* ***erklären****, da sei etwas rechtens und nicht strafwürdig, das doch gerade unrechtens und für die bürgerliche Gesellschaft hätte durchaus strafwürdig sein sollen*. (Demski 1993:56)

> *Von der Straße waren sie und ihr Mann in die Vorstellungen gestürzt, hatten Stück für Stück ungeduldig überprüft, ob es den Stürmen auch gewachsen war, ob es sie vielleicht sogar* ***erklärte****.* (Demski 1993:99)

> *Britomartis*
> *Kretische Göttin, von den Griechen mit Artemis gleichgesetzt; ihren Beinamen Diktynna (wohl nach dem Berg Dikte)* ***erklärte*** *man* ***damit****, daß Britomartis, von Minos verfolgt, ins Meer gesprungen und von Fischern mit einem Netz (gr. diktys) aufgefangen worden sei.* (Fink 1993:74)

‚äußern'

> *Er hat juristische Phantasie,* ***erklärte*** *Hardenberg und erläuterte hinter dem Rücken des Mannes dessen Beweggründe.* (Demski 1993:341)

> *In den Metamorphosen des Ovid ist der Streit um die Waffen des Achilleus breit behandelt (XII 620-XIII398); damit er sich in den Rahmen der Verwandlungsgeschichten fügt,* ***wird erklärt****, aus dem Blut des toten Aias sei eine purpurrote Blume entsprossen, auf deren Blütenblättern deutlich AI zu lesen sei*. (Fink 1993:23)

‚etw. als etw. bezeichnen'

> *Es war Eigenliebe gewesen, Angst davor, mit einem solchen Schritt ein Theaterstück zu wirklichem Leben zu* ***erklären***. (Demski 1993:157)

Wendung *den Krieg erklären*

> *Die Göttin verwandelte sie darauf in einen Kranich (gr. geranos) und ließ sie dem eigenen Volk* ***den Krieg erklären***. (Fink 1993:270)

‚Begründung finden'

> *Dasselbe teilt Livius* […] *mit, der außerdem schreibt, Acca Larentia sei eine Dirne (lupa) gewesen und auch so genannt worden. Wenn Romulus angeblich von einer Wölfin gesäugt wurde,* ***erkläre sich*** *das daraus, daß lupa eigentlich „Wölfin" bedeute*. (Fink 1993:13)

‚einen Heiratsantrag machen'

> *Das war der Moment, die Frau einzubeziehen, sich ihr zu* ***erklären****, sie zu gefährden. Ja, das auch. Aber sie war verstockt gewesen und konnte sich trotz ihrer Scham nicht von dem Räuber trennen. Den Bankräuber Toni hätte sie jederzeit politisch erklären können. Er gehörte zu ihnen, er hatte sich am weitesten ins Feindesland gewagt, er hatte eine proletarische Kindheit, eine Heimkarriere*. (Demski 1993:248)

Unser Korpus ist sehr klein, vermutlich zu klein, um als irgendwie repräsentativ gelten zu können, auch wenn es im Sprachlichen kaum eine Repräsentativität in mathematischem Sinn gibt. Allerdings lassen sich schon zwei Tendenzen erkennen:

- Die Bedeutungen ‚erläutern' und ‚äußern' werden am häufigsten verwendet; es überrascht daher, dass ‚äußern' von den gegenwartssprachlichen Wörterbüchern nicht bekannt zu sein scheint.
- Die Bedeutung ‚erläutern' begegnet hauptsächlich im narrativen Text, während ‚äußern' vom Lexikon verwendet wird. Es geht dabei immer darum, dass eine wichtige Instanz (Pfarrer, Autor, Orakel) etwas äußert.

Mein kleines Korpus habe ich durch die ‚Süddeutsche Zeitung' vom 2. Juni 2003 (dieses Datum wurde zufällig ausgewählt), erweitert. Die Präteritalform *erklärte* kommt in dieser Nummer insgesamt 12mal vor, davon 10mal im Ressort Nachrichten, einmal im Ressort Medien und einmal im Ressort Wirtschaft.

Im Ressort Medien wird das Verbum in der Bedeutung „erläutern" verwendet:

> *Am Anfang war die Geschichte der „Kinowelt" ein Märchen. In dem spielen die Brüder Michael und Rainer Kölmel, die als Mathematiker und Historiker ordentliche Jobs an der Uni hatten und nebenbei in Göttingen ein kleines Programmkino betrieben, die Hauptrollen. Film war ihre Leidenschaft. Irgendwann wollten sie Gregory's Girl, das sie sehr mochten, in viele Kinos bringen. Also fragten sie nach, wie das wohl so gehe. Dafür müssten sie die Rechte kaufen,* ***erklärte*** *man ihnen. Und so verscherbelten sie einen Polo und kauften für 20 000 Mark die Rechte.*

Das Verb *erklären* ist hier, wie im dtv-Korpus auch, dreiwertig, neben der Nominativergänzung kommen noch eine Dativ- und eine Akkusativergänzung vor.

In allen anderen Fällen hat das Verbum *erklären* die Bedeutung ‚äußern':

Ressort Nachrichten

> *Ein Dialog, so hat Suu Kyi in jüngster Zeit jedoch beklagt, ist bis heute nicht zustande gekommen. Es sei fraglich,* ***erklärte*** *die Oppositionsführerin mehrfach, ob die Junta überhaupt an Gesprächen mit den demokratischen Kräften interessiert sei.*

*Teheran **erklärte** jedoch, es befänden sich keine Führungsmitglieder der Organisation unter den Festgenommenen, deren Identität den iranischen Behörden zum Teil noch unbekannt sei.*

Ressort Wirtschaft

„Wir müssen alles tun, um solche Fälle zu vermeiden. Das Vertrauen in den Berufsstand der Wirtschaftsprüfer hat Schaden genommen, obwohl er zu 99,99 Prozent einwandfrei gearbeitet hat“, erklärte Wienand Schruff, Vorsitzender des Hauptfachausschusses beim Institut der Wirtschaftsprüfer (IDW) und Vorstandsmitglied der Prüfgesellschaf KPMG, im Gespräch mit der Süddeutschen Zeitung.

In dieser Lesart ist *erklären* zweiwertig, die zweite Ergänzung ist eine Propositionalergänzung, meistens realisiert als direkte oder indirekte Rede. Die Nominativergänzung wird durch eine Personenbezeichnung realisiert, die auf eine offizielle oder öffentliche Person referiert. Wir haben also ein ganz spezielles Skript vor uns; der berühmte Otto Normalbürger bekommt selten die Gelegenheit, etwas zu *erklären*; er kann nur etwas *sagen* oder *äußern*. Dieses Skript ist auch in den Belegen aus dem dtv-Korpus zu finden. Allerdings hat es den Anschein, dass Pressenachrichten diese Bedeutung aufgrund ihres speziellen Skripts bevorzugen.

Kleine oder kleinere Korpora haben – das zeigt unser Beispiel deutlich – durchaus ihren Zweck. Sie können vor allem bei hochfrequenten Phänomenen zumindest erste Tendenzhinweise geben. Und sie bieten Übersichtlichkeit, wo der Wust der Belege, die die großen Korpora ausgespuckt haben, den Betrachter schlicht und einfach zuschütten. Möglicherweise kann man die zahlreichen Belege zählen, doch darf das Zählen nicht mit der sprachwissenschaftlichen Interpretation des Befundes gleichgesetzt werden. Wir brauchen neben den Wortbelegen vor allem die Kontexte, aus denen heraus die Bedeutungsbeschreibung erfolgen muss. Bei der Korpusrecherche und -analyse müssen Textarten unterschieden werden können, da bestimmte Textarten bestimmte Verwendungsweisen bevorzugen.

Schließlich gibt die Korpusanalyse Hinweise auf Frequenzen von Bedeutungen und funktionale Belastungen einzelner Ausdrücke. Derartiges sollte z.B. bei der Erstellung von Wörterbuchartikeln berücksichtigt werden.

Beide Beispielfälle bestätigen das eingangs formulierte Postulat, dass jedes Projekt sein eigenes Korpus braucht. Dies ist natürlich überspitzt formuliert. Doch in Hinblick auf ein Referenzkorpus sollte sich die Forderung ergeben, dass die jeweiligen Spezialkorpora aus dem großen oder besser: aus dem sehr großen Referenzkorpus herausgenommen werden können, dass die möglichen Spezialkorpora im Referenzkorpus enthalten sind. Dies ist natürlich ein Ideal, das kaum zu erreichen ist, sodass auch Carmen Scherers (2006:27) eingangs zitierte Formulierung, dass ein Referenzkorpus „eine Sprache in ihrer Gesamtheit zu repräsentieren“ habe, nur als eine ideale Zielnorm anzusehen ist.

Zum Abschluss noch eine Überlegung: Streng genommen, sollten Korpora auch Metadaten und Annotationen enthalten. Metadaten sind unerlässlich, und wir haben von unseren zweisprachigen Würzburger Korpora sogar Bilder der Buchvorlagen gemacht, damit wir jederzeit das ursprüngliche Layout rekonstruieren können. Schwieriger wird es mit den Annotationen, da diese zu erzeugen ziemlich zeitaufwendig ist und der eigentlichen sprachwissenschaftlichen Arbeit vorausgehen müsste. Daher wird man bei Spezialkorpora häufig darauf verzichten (müssen). Für viele Fragestellung sind Annotationen auch nicht notwendig.

Wichtig bleibt, dass wir unsere Arbeit mit authentischen Sprachdaten machen und dass wir diese Daten immer auch kontextualisieren können. Die Sprache ist – dies kann man nicht oft genug wiederholen – nicht ein Mittel, schöne Strukturen zu generieren, sondern mit Hilfe eben dieser Strukturen Inhalte zu transportieren.

Literaturverzeichnis

Primärlitertur:

Demski, Eva (1993): *Scheintod*. München.

Fink, Gerhard (1993): *Who's who in der antiken Mythologie*. München.

Mythologie.

Süddeutsche Zeitung, 2.6.2003.

Süddeutsche Zeitung, 3.1.1994.

Sekundärliteratur:

Brendel, Bettina / Frisch, Regina / Moser, Stephan / Wolf, Norbert Richard (1997): *Wort- und Begriffsbildung in frühneuhochdeutscher Wissensliteratur. Substantivische Affixbildung.* Wiesbaden.

Dammers, Ulf / Hoffmann, Walter / Solms, Hans-Joachim (1988): *Grammatik des Frühneuhochdeutschen Bd. IV: Flexion der starken und schwachen Verben*. Heidelberg.

Duden (2011): *Duden. Deutsches Universalwörterbuch*. 7. Aufl. CD-ROM-Ausgabe. Mannheim.

Grimm, Christian (1989): Substantivische Affixbildung in wissensliterarischen Texten des Früneuhochdeutschen. In: Moser, Stephan / Wolf, Norbert Richard (Hrsg.): *Zur Wortbildung des Frühneuhochdeutschen. Ein Werkstattbericht*. Innsbruck, S. 65–86.

Habermann, Mechthild / Müller, Peter O. (1989): Verbale Wortbildung im Nürnberger Frühneuhochdeutschen am Beispiel er-. In: Moser, Stephan / Wolf, Norbert Richard (Hrsg.): *Zur Wortbildung des Frühneuhochdeutschen. Ein Werkstattbericht*. Innsbruck, S. 45–64.

Korhonen, Jarmo (1978): *Studien zu Dependenz, Valenz und Satzmodell* Tl. II. Bern; Frankfurt a. M.; Las Vegas.

Moser, Hans / Wolf, Norbert Richard (Hrsg.) (1989): *Zur Wortbildung im Frühneuhochdeutschen. Ein Werkstattbericht*. Innsbruck.

Prell, Heinz-Peter (1989): Zur Verbableitung bei Martin Luther. In: Moser, Hans / Wolf, Norbert Richard (Hrsg.): *Zur Wortbildung des Frühneuhochdeutschen. Ein Werkstattbericht.* Innsbruck, S. 39–44.

Prell, Heinz-Peter / Schebben-Schmidt, Marietheres (1996): *Die Verbableitung im Frühneuhochdeutschen.* Berlin; New York.

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg.

Wahrig (2012): *Wahrig. Deutsches Wörterbuch*. 9. Aufl. CD-ROM-Ausgabe. Gütersloh; München.

Wolf, Norbert Richard (2011): Wortbedeutung und Korpus. In: Lejsková, Alena / Valdrová, Jana (Hrsg.): *Die Grammatik, Semantik und Pragmatik des Wortes. Ihre Erforschung und Vermittlung.* Augsburg.

Internetquellen:

URL 1: www1.ids-mannheim.de/kl.html. [29.01.2017].

URL 2: http://www1.ids-mannheim.de/fileadmin/kl/dokumente/flyer-dereko-2015.pdf. [29.01.2017].

URL 3: www1.ids-mannheim.de/direktion/kl/projekte/korpora.html?L=0 [03.01.2018].