Sprachdaten sammeln und auswerten : die Korpuslinguistik am Institut für Deutsche Sprache (IDS)

Autor(en): **Anliker, Peter**

Objekttyp: **Article**

Zeitschrift: **Sprachspiegel : Zweimonatsschrift**

Band (Jahr): **70 (2014)**

Heft 5

PDF erstellt am: **13.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-422110

Sprachdaten sammeln und auswerten

Die Korpuslinguistik am Institut für Deutsche Sprache (IDS)

Von Peter Anliker[1]

Im Jahr 1964, also vor fünfzig Jahren, wurde in Mannheim das Institut für Deutsche Sprache gegründet (damals noch «Institut für deutsche Sprache» geschrieben). Mehrere Gründe führten dazu, dass sich von Anfang an Mitarbeitende des Instituts mit Korpuslinguistik befassten: Zum einen begann in dieser Zeit die grosse Ära der elektronischen Datenverarbeitung – es war also gewissermassen eine Frage der Zeit, bis sich auch Forschende der Germanistik dieses Hilfsmittels bedienen würden. Wichtiger aber waren wohl die politischen Beweggründe. Die Verwendung von Datenbanken und statistischen Auswertungen versprach die Betrachtung der Sprache als quasi «harte» Wissenschaft, die von belegbaren Fakten und nicht etwa von politischen oder ideologischen Wertungen bestimmt war.

Riegel gegen politische Vereinnahmung

Einerseits musste nämlich damals (in der Bundesrepublik Deutschland) immer noch die Zeit des Dritten Reichs verarbeitet werden, in der sich viele Germanisten dem Nationalsozialismus angenähert hatten. Nach der Trennung Deutschlands in die Bundesrepublik und die Deutsche Demokratische Republik ertönte aber auch der Vorwurf, in der DDR werde die Sprachwissenschaft wie die andern Geisteswissenschaften ideologisch (also im Sinne der herrschenden SED) betrieben. Die Abstützung auf Korpora, deren Inhalt aus «echten» Texten stammte (also aus literarischen Werken, Zeitungstexten, Gesetzessammlungen, später aber auch vermehrt Alltagstextprodukten wie etwa Gebrauchsanweisungen und Transkripten gesprochener Texte), sollte der Manipulation einen Riegel schieben.

1 Peter Anliker, Bern, ist Linguist. 1986/87 war er mit einem Stipendium als Gastforscher am IDS tätig und ist der Institution seither mit regelmässigen Besuchen freundschaftlich verbunden. Hauptberuflich ist er Journalist, heute bei der Zeitung «kontakt.sev». – *pan@die-politikerin.ch*

Korpusauswertung hat eine lange Tradition

Die Ursprünge der Korpuslinguistik allerdings reichen viel weiter zurück, auch wenn es die Bezeichnung ursprünglich noch nicht gab. Die Dokumentation und das Auswerten der (vorhandenen) Sprache spielten gerade in der Germanistik immer eine grosse Rolle: Die Materialbasis für das Grimmsche Wörterbuch etwa wurde erstellt, indem zahlreiche Beiträger literarische Texte nach den in ihnen verwendeten Wörtern durchforsteten und ihre Funde auf Karteikarten notierten. Das hübsche «Wörterbuch für Volksschulen» von Ludwig Wittgenstein entstand auf ähnliche Weise, indem Wittgenstein seine Schüler und Schülerinnen die von ihnen gekannten und gebrauchten Wörter sammeln und aufschreiben liess. Auch die Dudenredaktion stützte sich bis vor wenigen Jahrzehnten auf eine Kartei, deren Karten Wörter unter Angabe des Verwendungs- und Fundortes enthielten – heute kann man sich diese Arbeitsweise kaum mehr vorstellen.

Und natürlich wohnt diesem System ein nicht wettzumachender Fehler inne: Zwar werden alle (gefundenen) Wörter verzeichnet; welche davon aber schliesslich in einem Wörterbuch Aufnahme finden, muss ein Bearbeiter, eine Bearbeiterin entscheiden. Der «Faktor Mensch» kommt also immer wieder mit der «Objektivität» in Konflikt. Die Auswirkungen können wir alle selber nachprüfen, wenn wir im Duden ein bestimmtes Wort suchen – und nicht finden. Alte Wörter, die kaum mehr gebraucht werden, müssen «neuen» Wörtern Platz machen, die einen Aufschwung im Gebrauch erleben – wie frequent der Gebrauch eines neuen Wortes schon oder noch ist, dafür mussten sich die Bearbeiter auf ihr «Gefühl» verlassen, bevor ihnen Korpora und statistische Analyseinstrumente zur Verfügung standen. Und für komplexere Fragestellungen im grammatikalischen Bereich waren Belegstellensammlungen wie die oben genannten nicht zu gebrauchen.

Technisch schwierige und umständliche Anfänge

Doch mit dem Wunsch, über ein Korpus für die Recherche zu verfügen, beginnen die Probleme erst recht: In den Anfängen

der Korpuslinguistik mussten die Wörter auf Lochkarten erfasst und mechanisch eingelesen werden. Der zeitliche und materielle Aufwand eines solchen Verfahrens lässt sich leicht erahnen, noch viel mehr, wenn man daran denkt, dass die Wörter im Korpus annotiert werden müssen, damit sinnvoll mit ihnen gearbeitet werden kann. Und diese Annotationen müssen wiederum strikt kodifiziert werden.

Zudem waren am Anfang des Einsatzes der elektronischen Datenverarbeitung oder der linguistischen Datenverarbeitung, wie diese Methode auch genannt wird, vielfältige Fragen in Bezug auf die zu verwendende Software – oder, falls eine neue Software noch geschrieben werden musste, die zugrunde zu legende Programmiersprache – zu beantworten. Auch für jede Recherche im Korpus musste ein kleines Programm geschrieben und auf Lochkarten gestanzt werden. Die Rechner arbeiteten gemessen an den heutigen Geschwindigkeiten langsam, sie waren rar und teuer. In den ersten Jahren der Verwendung elektronisch gespeicherter Korpora hatten zudem nur die Informatiker das nötige Fachwissen, um aus den Beständen brauchbare Nachweise zu exzerpieren. Die Linguisten mussten sich also mit ihren Wünschen an die Informatiker wenden – eine zwar fruchtbare, aber nicht immer konflikt- und fehlerfreie Schnittstelle.

Die Anfänge der Mannheimer Korpuslinguistik

Im ersten, 1967 erschienenen Jahrbuch des Instituts für deutsche Sprache wird über die Anfänge der Mannheimer Korpuslinguistik so berichtet: «Die Abteilung ‹Dokumentation des heutigen Deutsch› soll ein vordringliches Bedürfnis erfüllen. In einem repräsentativen Querschnitt durch das heutige Schrifttum und aufgrund einer sorgsam vorbereiteten Programmierung soll zunächst die geschriebene deutsche Sprache von heute auf elektronischem Wege gespeichert werden, wobei Gesichtspunkte der Wortlehre und des Satzbaus berücksichtigt werden.»

Das erste Korpus hiess Mannheimer Korpus, später Korpus 1 genannt, als ein zweites dazukam. Es umfasst 293 Texte mit 2,2 Millionen Wörtern. Der Schwerpunkt der Texte lag auf der Literatur, enthalten

waren aber auch populärwissenschaftliche Bücher und journalistische Texte. Das weit kleinere Mannheimer Korpus 2 erfasste weitere Textsorten wie Broschüren oder Groschenromane. Die Auswahl der Textsorten erstaunt sicher nicht: Gerade wenn die Ressourcen begrenzt sind, wird man mit diesen zuerst das «Vorbildliche» abbilden. Freilich ist ein so gewonnenes Korpus dann aber nur begrenzt repräsentativ, es bildet nicht die ganze Sprache ab, sondern nur eine Auswahl an Textsorten und -schichten. Eine solche Auswahl genügt aber dem Anspruch nicht, das Korpus könne zeigen, «wie Sprache funktioniert».

Und darum geht es eigentlich beim korpusbasierten Arbeiten, in der deutschen Sprache wohl sogar ausgeprägter als in andern Sprachen: Während jene, die nur für den Alltagsgebrauch Deutsch schreiben oder reden, sich meist nicht scheuen, Werturteile über den eigenen, öfter aber über den fremden Sprachgebrauch zu fällen, fällt dies jenen, die sich auf wissenschaftlicher Basis mit der deutschen Sprache befassen, oft viel schwerer. Vorschriften werden von Korrektoraten gemacht, nicht aber von der Wissenschaft. Daher stammt diese merkwürdige Symbiose von Deskription und Präskription: Zuerst wird der Sprachgebrauch beschrieben, anschliessend das Beschriebene zur Norm erklärt.

Sprachrichtigkeit als Frage der Frequenz

Freilich tauchen damit neue Probleme auf. Denn nicht alles, was sprachlich möglich ist, ist richtig – richtig wird es erst durch die Verwendung. (Unter «Verwendung» verstehe ich hier das, was der späte Wittgenstein «Lebensform» nannte.) Als «richtig» werden dabei nur jene Sprachphänomene bewertet, die eine gewisse Frequenz erreichen – eine sprachliche Eintagsfliege kann verständlich und dennoch falsch sein. Ein Beispiel für das Gesagte mag das Wort «nichtsdestotrotz» sein, das bei seiner Entstehung Originalität, Witz und Verständlichkeit vereinigte und trotzdem falsch war. In der Zwischenzeit hat sich das Wort allerdings etabliert, sein Gebrauch ist frequent – und seit einiger Zeit ist es auch richtig, sein Gebrauch ist durch die Aufnahme in den Duden sanktioniert. Dass es dort als «ugs.» (umgangssprachlich) markiert ist, zeigt eben gerade eine gewisse Frequenz: «Dabei bedeutet die

Aufnahme in den Duden keinerlei Werturteil oder amtliche Anerkennung. Die Redaktion in Berlin (früher Mannheim) beurteilt nur, ob ein Wort ‹häufig› und ‹breit gestreut› vorkommt und ‹keine Eintagsfliege› ist», erläutert Daniel Goldstein im «Sprachspiegel» 4/2013, S. 122. Und eine solche Frequenz feststellen oder belegen könnte man nicht leichter als durch die (elektronische) Auswertung eines Korpus: «Grundlage ist ein Korpus elektronisch gespeicherter Texte», fährt Goldstein fort.

Freilich ist das genannte Beispiel eher trivial. Interessanter sind grammatische Phänomene, etwa im Bereich der Satzstellung, oder die Frage – um noch einmal ein Beispiel aus der Lexik zu nehmen –, welche Nomen mit welchen Verben oder mit welchen Adjektiven zusammen gebraucht werden (die sogenannte Ko-Okkurenz-Analyse). Das erste grosse Vorhaben, für das das Mannheimer Korpus ausgewertet wurde, war das Projekt «Grundstrukturen der deutschen Sprache». 17 Bände zu verschiedenen Aspekten dieser «Strukturen» sind zwischen 1971 und 1981 erschienen.

Methoden verfeinert, Datenbasis erweitert

In der Zwischenzeit entwickelte sich die Datenlinguistik am IDS weiter. Einerseits wurde experimentiert, welche Lösungen der Computer etwa im Bereich der automatisierten Kommunikation überhaupt zuliess. Auch spezielle «tools» wurden erarbeitet, etwa der Wortformengenerator «Molex», der alle Flexionsformen von mehr als 60 000 Wörtern generieren kann und es damit beispielsweise ermöglicht, bei einem entsprechenden Forschungsvorhaben nicht für alle Flexionsformen eine eigene Recherche durchführen zu müssen. Und natürlich ging parallel dazu der Ausbau des Korpus weiter: aus der damaligen Freiburger Aussenstelle des IDS kam beispielsweise ein Korpus gesprochener Sprache.

1992 zog das IDS in ein umgebautes ehemaliges Spital in der Mannheimer Innenstadt um, und im Zuge dieser Ortsveränderung wurde auch die Anschaffung eines neuen Rechners nötig. Der ganze Bereich der Korpusarbeit wurde neu konzipiert. In diesem Zusammenhang wurde das Projekt «Cosmas» gestartet, ausgeschrieben

«Corpus Search, Management and Analysis System» (siehe dazu auch «Sprachspiegel»-*Netztipps* in dieser Ausgabe S. 155 und zum IDS in Heft 4/2013, S. 122 f.). Dieses Projekt stellte die Korpusarbeit am IDS auf eine neue Basis. Es wurde eine neue Korpusplattform entwickelt, die verschiedenste korpuslinguistische Prinzipien formulierte. Damals wurde auch mit dem Aufbau des Deutschen Referenzkorpus (DeReKo) begonnen. Die vorhandenen Korpora wurden in die neue Plattform integriert und es wurden neue Texte akquiriert mit dem Ziel, ein «neuartiges, universelles Archiv des geschriebenen Deutsch» zu schaffen.

Freilich ein hoher Anspruch! Gegenwärtig hat das DeReKo einen Bestand von 24 Milliarden Textwörtern, etwa dreimal so viel wie im Vorjahr. Der Zuwachs geht auf zwei grosse Lizenzabschlüsse zurück, mit denen (digitale) Zeitungs- und Fachtexte mit mehr als 16 Milliarden Wörtern erworben wurden. Mit dieser gewaltigen Datenmenge wird der öffentliche Sprachgebrauch in Deutschland, Österreich und der Schweiz abgebildet. Auch die kontinuierliche Erweiterung des Korpus wurde durch die neuen Lizenzen forciert, gegenwärtig beträgt die Wachstumsrate 1,7 Milliarden Textwörter pro Jahr. Recherchen im DeReKo können via Cosmas nach einmaliger Registrierung (gratis) im Internet vorgenommen werden (*http://www.ids-mannheim.de/cosmas2*). Nach Angaben des IDS sind «über 30 000 Nutzer aus über 100 Ländern» registriert.

Annotiert und in Stichproben stratifiziert

Doch obschon die Menge der Texte und Wörter im DeReKo wie gesagt beträchtlich ist, ist es natürlich so, dass ein Korpus, so gross es auch sei, immer nur einen (kleinen) Ausschnitt des gesamten Sprachgebrauchs enthält. Um dennoch zu richtigen Aussagen zu kommen, um also aus dem im Korpus beobachteten Sprachgebrauch verallgemeinern zu können, muss das Korpus ausreichend repräsentativ sein für den Sprachgebrauch, den ein Forscher oder eine Forscherin untersuchen will. Um dies zu erreichen, arbeitet Cosmas mit einer sogenannten stratifizierten Stichprobenziehung. Es wird also nicht mit der gesamten Datenmenge gearbeitet, sondern nur mit einer Stichprobe, die aber nicht zufällig, sondern stratifiziert ist. Dabei wird der Sprachausschnitt,

der untersucht werden soll, in Hinblick auf verschiedene Merkmale in Gruppen eingeteilt, die Strata genannt werden. Solche Strata können beispielsweise Modus, Genus, Texttyp, Thema, Publikum oder Zeit sein.

Aus dem eben Gesagten geht schon hervor, dass es sich beim DeReKo nicht einfach um eine riesige, mit Texten gefüllte Tüte handelt. Vielmehr sind die Texte annotiert. Diese Annotationen umfassen die bibliografischen Daten wie Autor, Titel, Verlag, Erscheinungsdatum und Ort, aber zusätzlich auch die Entstehungszeit. Weiter umfassen die Annotationen deskriptiv-statistische Werte zu den Texten, textlinguistische Kategorisierungen, die historisch-politische Zuordnung und Angaben zur verwendeten Rechtschreibnorm. Dazu kommt die eigentliche linguistische Annotation, etwa zur Wortart. Das annotierte DeReKo umfasst heute hundertmal mehr Daten als die reinen Texte.

Der heutige Programmbereich Korpuslinguistik am IDS versteht sich einerseits als «Dienstleistungszentrum», indem er das DeReKo betreibt, ausbaut und pflegt. Mit berechtigtem Stolz zeigen die Wissenschafter des Instituts die Funktionsweise und die Möglichkeiten der Datenlinguistik und führen Fachleute in die Arbeit mit Cosmas II ein. Auf der andern Seite wird am IDS aber auch selbständig korpusbasiert linguistisch geforscht. Das Ziel der Forschungsarbeit ist es, neue Einsichten in die Strukturen, Gesetzmässigkeiten, Eigenschaften und Funktionen zu gewinnen. Für beide Aufgaben zentral ist es, einerseits die Methoden der Korpusanalyse und -erschliessung kontinuierlich weiterzuentwickeln, andererseits das Korpus auszubauen. Bereits wird an einer neuen Korpusanalyseplattform unter der Bezeichnung KorAP gearbeitet, die in einigen Jahren – nach anfänglichem Parallelbetrieb – Cosmas II ablösen soll und noch stärker auf spezifische Bedürfnisse bei den jeweiligen Forschungsarbeiten eingehen kann.

Jubiläumsband und Jahrbücher des Instituts für Deutsche Sprache

Der vorliegende Artikel stützt sich massgeblich auf Arbeiten zur Korpuslinguistik in: Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. IDS, Mannheim 2014.

Die Referate der IDS-Jahrestagung werden jeweils als Buch herausgegeben, zuletzt: Sprachverfall? Dynamik – Wandel – Variation. Hrsg. v. Plewnia, Albrecht / Witt, Andreas. De Gruyter Mouton, Berlin 2014. 379 Seiten, 99,95 Euro (auch als E-Book erhältlich).