FOLIA SCANDINAVICA
VOL. 1 POZNAŃ 1992

EIN TEXTKORPUS DER GESCHRIEBENEN DÄNISCHEN GEGENWARTSSPRACHE

HENNING BERGENHOLTZ

ABSTRACT. [A word corpus of contemporary written Danish] The paper deals with the main problems of word corpus studies. It includes a general survey of Danish word corpora compiled so far as well as it informs of the present stage of word corpus studies in Denmark.

1. TEXTKORPORA

In linguistischen Teildisziplinen wie der Sprachenforschung oder der historischen Linguistik ist es seit immer eine Selbstverständlichkeit gewesen, Texte als empirische Basis der Untersuchungen zugrunde zu legen. Auf dem Gebiet der Linguistik, die sich insbesondere mit der Morphologie und der Syntax der Gemeinsprache beschäftigt, ist dies seit Anfang der sechziger Jahre anders gewesen. Immer wieder wurde ein fast ritueller Kampf zwischen Skeptikern und Verfechtern einer Auswertung von Textkorpora der Sprachforschung geführt. Auf der einen Seite befanden sich die Korpusskeptiker, die die Auswertung von Korpora als "eine unnötige Zeremonie" ansahen, wie es Itkonen einmal formulierte. Diese Formulierung ist sogar positiv im Vergleich mit Chomskys berühmt gewordener Gleichnis: Ein Korpus auszuwerten, schrieb er, wäre ungefähr so vernünftig wie eine Beschränkung der Physik oder Biologie auf Filme, die von den Ereignissen handeln, die um uns herum in unserem Alltag stattfinden. Auf der anderen Seite vermitteln die Verfechter von Korpora teilweise einen habeas–corpus–Eindruck: Wenn Du ein Textkorpus hast, finden sich die nötigen Theorien von selber ein. Als eine Fortsetzung dieser theorie–naiven Einstellung sehe ich die m.E. unzweckmäßigen Bestrebungen, eine eigene neue Bindestrich–Linguistik, eine Korpus–Linguistik, zu begründen. Zu den jeweiligen Argumenten dieser Kontroversen nehmen Bergenholtz/Mugdan 1988 Stellung. In den letzten zwei–drei Jahren scheint diese Auseinandersetzung an Interesse zu verlieren. Der prinzipielle Wert von Korpora wird in geringerem Maße bestritten, und die Befürworter besinnen sich auf die Begrenzungen der vorliegenden Korpora.

Ein weiterer Diskussionspunkt ist es gewesen, ob und inwiefern ein repräsentatives Textkorpus erstellt werden kann. M.E. ist diese Diskussion durch Rieger[1] abschließend behandelt worden, der die Grundprinzipien der Statistik auf die Sprachwissenschaft überträgt. Eine Stichprobe kann dann als repräsentativ gelten, wenn sie hinsichtlich bestimmter Eigenschaften mit der Grundgesamtheit übereinstimmt, aus der sie stammt. Die Grundgesamtheit einer natürlichen Sprache ist jedoch nicht bekannt; dies gilt nicht nur für mündliche, sondern auch für schriftliche Texte. Der in der Statistik belegter Ausdruck "repräsentativ" ist daher nicht oder nur mit der Gefahr von Mißverständnissen verwendbar. In Anlehnung auf Bungarten[2] verwende ich selber den weniger anspruchsvollen Ausdruck "exemplarisch". Ein Korpus kann als exemplarisch gelten wenn man

1. statt der Menge aller Texte einer Sprache eine wohldefinierte Teilmenge als Grundgesamtheit wählt
2. eine plausibel scheinende hypothetische Verteilung bestimmter Textmerkmale annimmt und diese zugrunde legt.

Das erste Verfahren empfiehlt sich m.E. vor allem in den Teilen der Fachsprachenforschung, die sich mit technischen Fachsprachen beschäftigt. Das zweite Verfahren eignet sich m.E. am ehesten für die Erforschung der Gemeinsprache und für die nicht–technischen Fachsprachen.

In beiden Fällen spielt aber die Größe des Korpus eine wesentliche Rolle. Es gibt dabei große Differenzen, was den Umfang und die Länge der einzelnen Texte betrifft. Für die Gemeinsprache hat man oft ein Korpus mit einem Umgfang von einer Mio. Textwörtern als groß, ja gar als repräsentativ für eine bestimmte Sprache angesehen. Dies stimmt mit der Praxis überein, Korpora dieser Größe zusammenzustellen bzw. damit zu arbeiten, vgl. z.B. für das Deutsche das Limas–Korpus und für das Englische das Brown–Corpus, das LOB–Corpus und das Lancaster–Corpus. Dagegen haben Bergenholtz/Mugdan 1985[3] dafür argumentiert, daß erst ein Korpus mit fünf Mio. Textwörtern ausreichend wäre für eine lexikographische Darstellung der 2000 häufigsten Lexeme und Affixe der heutigen deutschen Gemeinsprache. Bei der Planung des nicht–realisierten großen interdisziplinären deutschen Wörterbuches wurde ein Korpus von 50 Mio. Textwörtern vorgesehen.[4] Hoffmann/Piotrowski[5] meinen unter Verweis bekannter Statistiker, daß ein repräsentatives Korpus einer natürlichen Sprache einen Umfang von zwischen 10^9 und $1.5x\ 10^{14}$ Textwörtern

---

[1] Rieger Burghardt, Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In: Bergenholtz/Schaeder, Empirische Textwissenschaft. Aufbau und Auswertung von Text–Corpora. Königstein/Ts.: Scriptor 1979.

[2] Bungarten Theo, Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. In: Bergenholtz/Schaeder 1979, S.28–51.

[3] Bergenholtz Henning/ Mugdan Joachim, Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität.In:Computional Linguistics Ein internationales Handbuch computerunterstützter Sprachforschungund ihrer Anwendung. Hrsg. von Istvan Batori u.a., Berlin New York:de Gruyter (im Druck).

[4] Mentrup Wolfgang, Überlegungen zur Zusammenstellung und Verwendung eines Korpus für ein großes interdisziplinäres Wörterbuch der deutschschen Sprache. In:Bergenholtz/Schaeder 1979, S.182–203.

[5] Hoffmann L.,/Piotrowski R.G., Beiträge zur Sprachstatistik. Leipzig VEB Enzyklopädie 1979.

haben müßte. Abgesehen von dem problematischen Gebrauch des Terminus "repräsentativ" ist dagegen einzuwenden, daß ein Korpus dieser Größenordnung nicht von den heutigen Computern bewältigt werden könnte. Dieser Einwand ist auch geltend zu machen gegenüber Bahr 1987,[6] der mit Verweisen auf Hoffmann/Piotrowski 1979 vorsieht, ein Korpus mit etwa 500 Mio.Textwörtern für ein neues, historisches deutsches Wörterbuch zusammenzustellen. Demgegenüber mutet wenig an, was Politov 1987[7] vorschlägt: Der Umfang einer Gesamtstichprobe für eine fachsprachliche Gesamtstichprobe solle zwischen 20.000 und 35.000 Textwörtern liegen. Politov schlägt weiterhin vor, jede Teilstichprobe etwa 200 Textwörter groß werden zu lassen.

Hiermit ist ein weiterer Problembereich angesprochen: Wenn der Umfang der einzelnen Texte in einem Korpus einer gegebenen Größe sehr groß wird, wird die Zahl der Texte geringer werden müssen. Die Gefahr, ein Korpus zu erhalten, daß kaum das Prädikat exemplarisch für eine bestimmte Sprache verdient,wächst entsprechend. Wenn aber sehr kleine Teiltexte in das Korpus eingehen, wird das Korpus nur für wortstatistische Untersuchungen herangezogen werden können. Textanalysen wären kaum durchzuführen, auch für lexikographische Zwecke werden Teiltexte mit einer Länge von z.B. 200 Textwörtern nur bedingt unbrauchbar (alle Anfänge und Schlußpassagen der Teiltexte wären nicht oder schlecht interpretierbar).

Bis vor einigen Jahren wurden Textkorpora rein manuell von Schreibkräften aufgenommen. Die anschließende Korrektur war sehr zeitaufwendig, und selbst nach einer zweifachen Korrektur mußte eine beträchtliche Fehlerquote (bis zu 1%) hingenommen werden. Mit dem Einsatz von optischen Lesegeräten ist eine erhebliche Vereinfachung möglich geworden. Mit Hilfe von Korrekturprogrammen können viele, wenn auch nicht alle Fehler korrigiert werden. In dem hier vorgestellten dänischen Textkorpus war die Fehlerqoute nach dem Scanning, aber vor der automatischen Korrektur geringer als eine Promille.

2. BISHERIGE DÄNISCHE TEXTKORPORA

Neben vielen kleinen, aber auch kaum allgemein zugänglichen dänischen Korpora gibt es die von Maegaard/Ruus zusammengestellten Korpora. Für folgende fünf Textgattungen wurde je ein eigenes Korpus zusammengestellt: Kinderbücher, Romane, Zeitungen, Wochenzeitschriften und populäre Fachzeitschriften. Häufigkeitsangaben zu den einzelnen Korpora, aber nicht zu dem Gesamtkorpus liegen in Buchform vor. Jedes Korpus besteht aus 1000 Textteilen mit jeweils 250 laufenden Wörtern, d.h. etwa eine Seite pro Korpustext und insgesamt 250 000 Textwörter pro Korpus. Bei längeren Textstücken sei die Gefahr gegeben, daß das Thema eines bestimmten Textes zu zufälligen Häufigkeitserscheinungen führen würde.[8] Die

---

[6] Bahr Joachim, Entwurf eines historischen Wortschatzarchivs. In: Zeitschrift für germanistische Linguistik. Deutsche Sprache in Gegenwart und Geschichte 15,1987, S.141–168.

[7] Politov Stefan, Zur Entwicklung der statistischen Fachsprachenlexikforschung. In: Fachsprache 9, 1987, S.149–166.

[8] Maegaard Bente/Ruus Hanne, Hyppige Ord i Danske Romaner. København: Gyldendal 1981, S.7.

Auswahl der Texte ist im Prinzip aufgrund von lesesoziologischen Untersuchungen vorgenommen worden. Bei Zeitungen, Wochenzeitschriften und Fachzeitschriften ist von den jeweiligen Auflagezahlen der Jahre 1970– 1974 ausgegangen worden: je höher die Zahl der Textproben dieser Zeitung bzw. Zeitschrift[9]. Bei den Kinderbüchern wurde (wahrscheinlich aufgrund fehlender lesesoziologischer Untersuchungen) vom folgenden Prinzip ausgegangen: Nur die dänischen Verfasser, die in der Zeit 1970–1974 mehr als drei Werke herausgebracht hatten, kamen für die Auswahl in Frage. Ausgewählt wurden neben Bilderbüchern für die Kleinsten, auch Erstlesebücher für Schulanfänger sowie Romane für größere Kinder, dabei sowohl neuaufgelegte Werke wie Neuauflagen. Bei der Auswahl der Romantextproben finden sich Texte der "am meisten gelesenen dänischen Schriftsteller" der Jahre 1970–1974[10]. Die Liste dieser Schriftsteller umfaßt 20 Namen, die ein weites Spektrum von Steen Steensen Blicher (1782–1848), H.C.Andersen (1805–1875), Herman Bang (1875–1912), Martin A. Hansen (1909–1955) bis zu "heutigen" Schriftstellern wie Klaus Rifbjerg und Anders Bodelsen bieten.

Dieses Korpus geht im wesentlichen vom Prinzip der Textrezeption aus, wenn auch dieses Prinzip bei der Auswahl der Roman– und Kinderbuchtextproben nicht strikt beachtet wurde, da hier nicht die Höhe der Auflagen die Selektion steuerte. Auch wäre zu bedenken gewesen, ob die Auflagenhöhe bei dem Grundprinzip der Verbreitung allein ausschlaggebend sein sollte. M.E. werden z.B. gewisse Wochenzeitschriften in höherem Maße als andere an Freunde weitergegeben.

Interessanter als diese methodischen Überlegungen halte ich folgende prinzipielle Problemstellungen:

1. Homogenität eines Korpus
2. Zeitliche Streuung der Texte
3. Zahl der Texte
4. Umfang der Teiltexte

Zur ersten Problemstellung äußern sich Maegaard/Ruus 1980[11] sehr entschieden. Sie gehen von der Forderung aus, ein Textkorpus müsse homogen sein, andernfalls würden die statistischen Ergebnisse nur für einen Text und nicht für die Textsorte insgesamt gelten können. Der Eckterminus ist dabei "homogen". Wie gleichartig müssen die Elemente einer Menge sein, die als homogen bezeichnet werden soll? Wir haben hier folgendes Dilemma: Je enger die Textssorteneingrenzung, je gleichartiger die Textsammlung, aber auch umso spezieller die Aussagen, die Untersuchungen an dieser Textsammlung ergeben. Wenn man umgekehrt von Auswahl der Texte ausgeht, die von einer Mehrzahl von Erwachsenen einer Sprache gelesen wird, wird die Sammlung weniger homogen, aber für die betreffende "Gemeinsprache" breiter verallgemeinbar.

---

[9] Maegaard Bente/Ruus Hanne, Hyppige Ord i Danske Aviser, Ugeblade og Fagblade, 2.Bde. København 1986, II, S.6–9.

[10] Maegaard Bente/Ruus Hanne, Hyppige Ord i Danske Romaner. København: Gyldendal 1981, S.5f.

[11] Maegaard Bente/Ruus Hanne, Danske almindelige ord rangfrekvenslister og deres brug. In: SALM 1, 1980, S.8.

Ich halte die einzelnen Textsorten in den besprochenen fünf Textkorpora für unbedingt homogen: Es bestehen z.B. große Unterschiede im Roman–Korpus zwischen Blicher aus der ersten Hälfte des 19. und Rifbjerg aus der zweiten Hälfte des 20. Jahrhunderts – auch aus der Sicht des Lesers. Dasselbe gilt im Kinderbuch–Korpus zwischen einem Erstlesebuch und einem Roman für größere Schulkinder. Auch finden sich wesentliche Unterschiede zwischen einer Boulevard–Zeitung aus Kopenhagen und einer bürgerlichen Provinzzeitung.

Was die Zahl der Texte und die Größe der Teiltexte, die in das Korpus eingehen, betrifft, wäre es in der Tat vorteilhaft mit einer möglichst großen Zahl von möglichst langen Teiltexten. Kleine Teiltexte mit 250 laufenden Wörtern sind nur für wortstatistische Untersuchungen zu verwenden, aber unbrauchbar für alle Formen der Textanalyse. Die von Hoffmann/Piotrowski angeführte Größenordnung für ein statisch gesehen ausreichend großes Korpus könnte sicher sowohl die Forderung einer breiten Textstreuung als auch die der Länge der Texte erfüllen. Bei dem jetzigen Stand der technischen Entwicklung müssen solche Forderungen als utopisch abgewiesen werden. Statt dessen sind pragmatische Lösungen gefragt.

3. EIN KORPUS DER DÄNISCHEN SCHRIFTLICHEN GEMEINSPRACHE

An der Wirtschaftsuniversität in Århus werden insbesondere Forschungen zum Thema Fachsprache durchgeführt. In allen Teilen dieser Forschung sind Kenntnisse der Gemeinsprache erforderlich, dies gilt besonders für die Wirtschaftssprache und die juristische Sprache. Mich selber beschäftigen z.Z. Fragen der Lexikographie, sowohl allgemeine ein– und zweisprachige metalexikographische Untersuchungen als auch die zweisprachige Fachlexikographie. In dem Zusammenhang werden zur Zeit an der Wirtschaftsuniversität Kopenhagen in Zusammenarbeit mit der Wirtschaftuniversität Århus auch drei Korpora zum Thema Vertragsrecht zusammengestellt, und zwar ein dänisches, ein englisches und ein französisches mit jeweils einer Mio. Textwörtern[12]. Gleichzeitig ist ein dänisches Textkorpus mit gemeinsprachlichen Texten erstellt worden. Die Prinzipien der Zusammenstellung und der Auswahl sind in Zusammenarbeit mit Finn Frandsen, Ole Lauridsen und Karen M. Lauridsen (alle Århus) erfolgt. Die konkrete Auswahlarbeit ist unter meiner Leitung von Lisbeth Boel, Louise Uggerhøj und Richard Almind ausgeführt worden, das ext-scanning ist innerhalb von einem Monat von einer privaten Firma ausgeführt worden.

Es ist ein Korpus zusammengestellt worden, das wie gesagt 1 Mio. Textwörter aus Originaltexten des Jahres 1987 enthält, d.h., daß keine Übersetzungen und keine Neuauflagen als Texte ins Korpus aufgenommen werden. Anders als bei den Korpora von Maegaard und Ruus sind wir somit von dem Prinzip der Textproduktion (und nicht Rezeption) ausgegangen. Das Korpus enthält keine Texte, die gezielt für Kinder verfaßt wurden. Es besteht aus drei Textarten:

---

[12] Dyrberg Gunhild / Faber Dorrit / Hansen Steffen Leo / Tournay Joan, Etablering af et juridisk tekstkorpus. Hermes 1, 1988 (im Druck).

1. Romane und Novellen (50% aller Texte)
2. Zeitungen (25% aller Texte)
3. Wochenzeitschriften (25% aller Texte)

Jeder Korpustext besteht bei Romanen und Novellen aus mindestens einem Kapitel, oder, wenn keine Kapiteleinteilung vorliegt, aus einem Textteil einer möglichst geschlossenen Einheit. Erstrebt wurde eine Textlänge von etwa 5.000 laufenden Wörtern, die in der Praxis nur geringfügig etwas unter– bzw. überschritten wurde. Es finden sich gut zehn Texte mit Trivialliteratur (Kioskromane), die restlichen knapp 90 Titel umfassen den größten Teil der 1987 auf Dänisch veröffentlichten neuen Romane. Aus den bekanntesten Zeitungen und Wochenzeitschriften wurden drei bis fünf Exemplare pro Zeitung bzw. Zeitschrift Texte und daraus jeweils etwa 5.000 laufende Wörter pro Exemplar ausgewählt.

Das Korpus für 1987 liegt in Form von 3.5" Apple–McIntosch–Disketten oder IBM 5.25" Disketten vor. Wissenschaftlern an Forschungsinstitutionen in Dänemark, aber auch in begrenztem Umfang im Ausland, kann das Korpus unter Beachtung folgender Bedingungen zur Verfügung gestellt werden:

1.das Korpus wird in der Forschung verwendet
2.aber unter keinen Umständen kommerziell genutzt
3.das Korpus darf nicht kopiert werden.

Die Disketten enthalten neben den Klartexten ein von John Bergenholtz geschriebenes Konkordanzprogramm, das nach Wahl eine links– oder rechtsalphabetische Satzkonkordanz generieren kann. Außerdem enthält die erste Diskette eine Beschreibung dieses Programms sowie bibliographische Hinweise zu Arbeiten zum oder mit dem Korpus.

Für die Jahre 1988, 1989, 1990 und 1991 werden weitere Korpora der dänischen Gemeinsprache zusammengestellt, die als Gesamtkorpus der Jahre 1987–1991 vorliegen werden. Nicht – oder noch nicht– geplant sind Textaufnahmen für die folgenden Jahre, was langfristig auch für diachrone Forschungen interessant werden könnte.

Das bereits vorliegende hier beschriebene dänische Korpus, aber auch die bis Ende 1988 fertiggestellten drei Korpora mit englischen, französischen und dänischen Texten zum Thema Vertragsrecht sind erhältlich bei

Prof. Dr. Henning Bergenholtz
Wirtschaftsuniversität Århus
Fuglesangsallé 4
DK – 8210 Århus V

(Eingegangen Juni 1988)