Ole Lauridsen / Theis Riiber / Henning Søndergaard

Erstellung eines dänischen und eines deutschen Textkorpus — Fachsprache Gentechnik

1. Um die Kommunikation in der Wirtschaft zu stärken — nicht zuletzt im Hinblick auf den Europäischen Binnenmarkt, beschloß die dänische Forschungsgemeinschaft 1987, die fachsprachliche Forschung zu fördern, und aufgrund eines von den Wirtschaftsuniversitäten ausgearbeiteten Forschungsprogramms wurde eine Reihe von entsprechenden Forschungsprojekten in Gang gesetzt — darunter die Erstellung maschinenlesbarer fachsprachlicher Textkorpora, eine Initiative, die um so wichtiger ist, als vor 1987 nur fünf derartige Textkorpora vorlagen: zwei deutschsprachige, das *Adelaide-Korpus DURF* und das *LIMAS Kfz-Korpus* sowie die drei an der Wirtschaftsuniversität Kopenhagen erarbeiteten *JUR-Korpora* in jeweils dänischer, englischer und spanischer Sprache.

Im Herbst 1987 leiteten Forscher der Wirtschaftsuniversitäten in Aarhus und Kopenhagen die Arbeit an der Erstellung von drei vertragsrechtlichen Korpora in den Sprachen Dänisch, Englisch und Französisch ein, und im Sommer 1988 beschloß die Forschungsgemeinschaft, weitere drei Korpora erstellen zu lassen, und zwar auf dem Gebiet der Gentechnik und in den Sprachen Dänisch, Deutsch und Spanisch. An diesem Projekt waren alle Wirtschaftsuniversitäten Dänemarks beteiligt; im folgenden sei indessen nur auf das dänische und das deutsche Korpus eingegangen.

Die Forschungsgemeinschaft legte in ihrer Ausschreibung den Interessenten nahe, die Korpora so zu gestalten, daß die darin enthaltene Fachsprache teils von wissenschaftlicher, technischer, parlamentarischer und journalistischer Provenienz ist, teils eine dementsprechende Funktion hat. Diese Formulierung bezeichnet eine Verdichtung und zugleich eine Entschärfung der ursprünglichen, von dem ehemaligen Mitglied der Forschungsgemeinschaft Henning Fonsmark aufgestellten Kriterien, nach denen die Korpora Texte enthalten sollten,

- die die Fachsprache in ihrer spezifischsten Form, d.h. die eigentlich wissenschaftliche Sprache widerspiegeln
- in denen sich die erste Popularisierung der Fachsprache manifestiert, wie z.B. Lehrbücher

- die, die durch die Implementierung der wissenschaftlichen Ergebnisse in Wirtschaft und Verwaltung verwendete Fachsprache zeigen
- die die Fachsprache wiedergeben, wie sie von den Parlamentariern und von der Presse gehandhabt wird.

Die Erfüllung der erwähnten Kriterien erfordert einen vertikalen Zugang zu einer Fachsprache und stellt somit einen Bruch dar mit der Praxis früherer dänischer Korpora (der JUR-Korpora und der vertragsrechtlichen), denn diese beschränkten sich in der Textaufnahme alle auf das eigentliche Expertenniveau und hatten damit ein horizontales Grundkonzept. Der vertikale Zugang sollte Untersuchungen der Änderungen ermöglichen, die — wie es wieder Henning Fonsmark im Namen der Forschungsgemeinschaft formulierte — die Fachsprache notwendigerweise erfahren muß, wenn ihre Botschaft einem immer breiteren, immer weniger spezialisierten Kreis von Interessenten vermittelt werden soll, ohne daß die Fachsprache damit ihren fachsprachlichen Charakter ganz verliert; diese Formulierung geht, wie man erkennt, in bezug auf die Popularisierung der Fachsprache ein bißchen weiter als die oben zitierten und bezieht letztendlich auch Leserbriefe, Feuilletons etc. ein, und da die Gruppe es wesentlich fand, auch Untersuchungen des (Fach)Sprachgebrauchs der Laien zu ermöglichen, beschloß sie, auch solche Texte mit aufzunehmen.

Daß Gentechnik als Fachgebiet der Korpora ausgewählt wurde, beruht ohne Zweifel auf dem herrschenden Trend. In mancher Hinsicht wären Textsammlungen aus anderen Fachgebieten, etwa dem merkantilen oder dem energietechnischen, aus Unterrichts- und Forschungsgründen wünschenswerter; andererseits ist es nicht unwesentlich, daß neue Fachsprachen möglichst schnell der Linguistik zugänglich gemacht werden, und außerdem ist die Korpuserstellung an sich von großem prinzipiellem Interesse.

Es mag vielleicht verwundern, daß die Fremdsprachenwahl auf gerade Deutsch und Spanisch fiel, da die Sprache der Gentechnik par excellence Englisch ist. Dies läßt sich aber wahrscheinlich damit begründen, daß die englische Fachsprache schon im vertragsrechtlichen Korpusprojekt berücksichtigt worden war und die Forschungsgemeinschaft im Forschungsprogramm keine der traditionellen Fremdsprachen der Wirtschaftsuniversitäten vernachlässigen wollte.

Der Umfang der Korpora wurde von seiten der Forschungsgemeinschaft auf je 1 Mio. Wortstellen festgelegt, so wie es allmählich Tradition geworden ist. Diese Zahl bestimmt wieder einmal nur der Trend — sie ist

nicht wissenschaftlich begründet, sondern völlig arbiträr; es hat sich jedoch in der Praxis herausgestellt, daß Korpora dieser Größe unmittelbar handhabbar sind und außerdem eine angemessene Grundlage für jedenfalls grammatisch-syntaktische Untersuchungen bilden.

Der Zweck der Korpora wurde weder in der Ausschreibung noch in den erwähnten Voraussetzungen angegeben, aber bei der Erstellung früherer Korpora in Dänemark und z.T. ebenfalls im Ausland war man generell bestrebt, daß das einzelne Korpus die Durchführung von Diskursanalysen, grammatisch-syntaktischen, lexikografischen sowie terminologischen Untersuchungen ermöglichen sollte; dies faßten wir deshalb als eine Forderung auch an unsere Korpora auf.

Ein Zeitrahmen für die aufzunehmenden Texte war ebenfalls nicht von vornherein festgelegt; es erschien uns aber natürlich, Texte einer etwa 10jährigen Periode — 1978-1989 — aufzunehmen, d.h. Texte von der Zeit an, wo Gentechnik allmählich ins Blickfeld einer breiteren Öffentlichkeit rückte, bis zum Abschluß des Projekts 1989.

Dies waren insgesamt die Voraussetzungen, von denen aus wir unsere Arbeit anfangen konnten — ein, wie aus den Ausführungen hervorgeht, angemessener Rahmen ohne, sagen wir, rigorose Einschränkungen und strikte Detailforderungen.

2. Eine erste Voraussetzung für die Erschließung einer Fachsprache, in diesem Fall also die der Gentechnik, bilden einmal eine Definition des Begriffs Gentechnik, zum anderen eine Abgrenzung des Fachgebiets gegen andere. Wir mußten aber dabei konstatieren, daß die Fachwelt — national sowie international — weder eine eigene Definition noch eine Fachsystematisierung ausgearbeitet hat, und auch das dänische Gesetz über Umwelt und Gentechnik zeigte sich alles andere als brauchbar; erstens ist es terminologisch recht vage und liefert nur wenige Schlüsselbegriffe zum Gebiet, zweitens — was noch schlimmer ist — sind gewisse Einzeldefinitionen unpräzise und stellenweise sogar unkorrekt, wie aus einem Gutachten des sogenannten "Gentechnologischen Rates" hervorgeht. Da uns auch die einschlägige ausländische Literatur im Stich ließ, entschieden wir uns dafür, bei den Literaturrecherchen — die erste Stufe unserer Arbeit — den Text selbst, in der Praxis den Absender als ausschlaggebenden Faktor gelten zu lassen: behauptete der Absender, sich über die Gentechnik zu äußern, d.h. machte der Absender unter irgend einem Aspekt Aussagen zu diesem Thema, gehörte der Text in unsere Interessensphäre, auch wenn seine Auffassung von Gentechnik keiner genauen wissenschaftlichen Prüfung standhalten würde.

Mangels definitionsgebundener Schlüsselbegriffe kreisten wir jetzt aufgrund von Texten, deren Autoren sich nach eigener Auffassung über die Gentechnik äußerten, eine Reihe von Suchwörtern ein, die eine Berücksichtigung aller relevanten Themen sicherten, wie wir uns auch von Fachleuten beraten ließen. Es zeigte sich dabei notwendig, nicht nur auf den Sprachgebrauch 1988/1989 zu fokussieren, sondern auch den früherer Jahre mit einzubeziehen, denn die Terminologie hat sich im Laufe der Zeit in gewissen Punkten geändert; so sprach man z.B. in Dänemark vor 7-8 Jahren von "genetisk manipulation" und "genmanipulation", jetzt wohl ausschließlich von "gensplejsning".

Bei der Literaturrecherche wurden vornehmlich Datenbanken, dänische wie ausländische, benutzt. Außerdem beschafften wir verschiedenes Textmaterial durch Anfragen bei Firmen, bei dem Presse- und Informationsdienst des Europaparlaments und der Kommission, durch Vermittlung des dänischen Patentamts und der Gewerbeaufsicht und nicht zuletzt durch die ständige Kontrolle von Hinweisen in der nach und nach eingegangenen Literatur.

3. Jedes geordnete Korpus ist seinem Auswahlkriterium und seinen sonstigen Konstruktionsprinzipien nach zu beurteilen — und für uns war in dieser Beziehung das Hauptanliegen von vornherein die Forderung nach möglichst großer Objektivität; vgl. hierzu das eben skizzierte Verfahren bei den Literaturrecherchen und der Textauswahl. Recht früh zeichnete sich ein Modell ab, das es uns ermöglichte, bei der Organisation der Korpora den Wünschen der Forschungsgemeinschaft zu entsprechen, ohne jeden einzelnen Text auf strukturelle, grammatisch-syntaktische usw. Merkmale subjektiv analysieren zu müssen.

Das Ergebnis unserer Überlegungen war ein stark vereinfachtes und konsequent verwendbares Klassifikationsprinzip nach dem Absender/Empfänger-Verhältnis; nicht als Auswahlkriterium (bei der Auswahl der Texte galten bei uns ja die Suchwörter), sondern als Einteilungskriterium — d.h. als ein äußeres objektives Kriterium, das keine vorausgehende linguistische Analyse erfordert. Natürlich schwingt die Subjektivität hier mit — dies aber nur bei der Festlegung des Kriteriums.

Ausgehend vom Status des Absenders als Laie oder Fachmann, wobei mit Fachmann jede Person gemeint ist, die sich professionell mit Gentechnik beschäftigt und/oder eine entsprechende fachliche Ausbildung hat, sowie der von ihm getroffenen Festlegung des Empfängerstatus stellten wir drei Kommunikationsebenen auf:

I.	Fachmann	↔	Fachmann
II.	Fachmann	↔	Laie
III.	Laie	↔	Laie

Daraus ergeben sich vier Kommunikationsrichtungen:

I.	Fachmann	→	Fachmann
IIa.	Fachmann	→	Laie
IIb.	Laie	→	Fachmann
III.	Laie	→	Laie

Es dürfte hieraus hervorgehen, daß die gewählte Struktur eine Aufhebung der klassischen Unterscheidung zwischen Fachsprache und Gemeinsprache bedeutet, weil in den Texten unserer Korpora das Sachgebiet nicht ausschließlich von Experten behandelt wird — und eben die sprachliche Handhabung eines Fachgebiets allein durch Experten gilt ja sonst als das basale Element der Fachsprachendefinition. Da unsere Korpora themenspezifische Texte von sowohl Experten als auch Laien enthalten, liegt hier eher vor, was wir als *Sachsprache* bezeichnen möchten; dieser neue Begriff läßt sich somit als eine Durchschnittsmenge von Fachsprache im klassischen und Fachsprache im weitesten Sinne des Wortes definieren. Nur aus praktischen Gründen haben wir im vorstehenden den herkömmlichen Terminus benutzt.

Die Bestimmung von Absender- und Empfängerstatus erfolgte aufgrund von Angaben auf Buchhüllen, in Vorworten, Bibliographien etc. In Zweifelsfällen wendeten wir uns an den Absender, an die Verlage oder die Redaktionen; nur Texte, bei denen die Statusfrage geklärt werden konnte, wurden aufgenommen. Wo ein Kollektiv von Fachleuten und Laien als Absender auftritt, ist der Text nach Absenderstatus *Fachmann* klassifiziert worden; zerfällt ein Text aber in eindeutige Experten- und Laienteile (etwa in Interviews), ist eine Doppelklassifizierung vorgenommen worden, und im bibliographischen Record ist angegeben, welche Teile welchem Niveau angehören.

Der vom Absender angegebene Status des Empfängers war ebenfalls für die Kategorisierung relevant: meint ein Absender, daß die Empfängergruppe aus sowohl Fachleuten als Laien besteht, ist eine entsprechende Doppelklassifizierung vorgenommen worden — egal, ob wir den Text als für Laien unverständlich hielten.

Bei dieser Kategorisierung verzichteten wir also völlig auf eine subjektive Beurteilung der Texte und meinen, dadurch maximale Objektivität erreicht zu haben.

Eine quantitativ gleichmäßige Verteilung der Texte auf die vier Kommunikationsrichtungen wurde angestrebt, wodurch jede der vier Kategorien im Prinzip 250.000 Wörter enthalten sollte. Trotz einer intensiven Literaturrecherche konnte dies Ziel nicht erreicht werden, und deshalb wurde das Wortdefizit in einer Kategorie durch Erhöhung der Wortanzahl in den anderen ausgeglichen, um die vorgeschriebene 1 Mio. Wörter zu erreichen. Doppelklassifizierte Texte wurden nach dem quantitativ dominierenden Textanteil kategorisiert, jedoch nicht wenn die Doppelklassifizierung aufgrund des Empfängerstatus erfolgte — in dem Fall wurden die Wortstellen auf die beiden Kategorien gleichmäßig verteilt.

Die prozentuale Verteilung von Wortstellen auf die vier Kategorien ist in beiden Korpora auch nicht die gleiche. Z.B. kommt Kategorie I (Fachmann —> Fachmann) im deutschen Korpus häufiger vor als im dänischen; dies muß der Benutzer bei kontrastiven und komparativen Studien selbst berücksichtigen. (Näheres s. Anhang 1).

Der Umfang der Textexzerpte wurde auf 5000 Wortstellen im laufenden Text festgelegt, er kann jedoch diese Grenze um 10% übersteigen, wenn das Exzerpt dadurch ein natürliches Ganzes ausmacht. Die Wahl dieses Umfangs ist weder wissenschaftlich begründet noch wissenschaftlich begründbar, es leuchtet aber unmittelbar ein, daß das Exzerpt weder zu begrenzt sein darf, da textlinguistische Untersuchungen dann nur schwer durchführbar wären, noch zu umfangreich, da die Zahl der Exzerpte dadurch zu gering wäre. In besonderen Fällen wichen wir aber von der erwähnten Grenze ab, u.a. bei der Kategorie IIb (Laie (meistens Politiker) —> Fachmann), um hier eine zu große Unterrepräsentanz zu vermeiden; dementsprechend ist z.B. das dänische Gesetz über Umwelt und Gentechnik sowie die entsprechende Gesetzesvorlage voll und ganz mit einbezogen.

Die Auswahl der Exzerpte aus langen Texten erfolgte im Prinzip vom jeweils einleitenden, mittleren und abschließenden Teil, um ein Übergewicht von einführenden bzw. abschließenden Abschnitten zu vermeiden und eine eventuell steigende Komplexität zu berücksichtigen.

4. Die Wahl des basalen, von den vier Kommunikationsrichtungen bestimmten Einteilungskriteriums ist wie schon betont an sich subjektiv, die praktische Anwendung des Kriteriums aber ist und bleibt objektiv. Wir haben indessen auch subjektive Kategorisierungen vorgenommen, aber nicht als eine sekundäre Grundlage für die Konstruktion der Korpora, sondern als Orientierungshilfe für den Benutzer bei dessen Recherchen in dem EDV- gespeicherten bibliographischen Record — es handelt

sich um Auskünfte über Aspekt und Publikationsmedien.

Die Kategorisierung nach Aspekt basiert auf der Sicht des Absenders, wenn er sich zur Gentechnik äußert. Die ausgewählten Aspekte, für die übrigens "Dansk Artikelindeks" als Vorbild diente, sind die folgenden:

Ethik
Generelles
Genetik
Industrie
Jura
Krankheit
Lebensmittel
Mensch
Militär
Pflanzen
Pharmazie
Politik
Sicherheit
Tiere
Umwelt
Unterricht
Wirtschaft

Beschäftigt sich ein Autor z.B. mit der Freisetzung genmanipulierter Pflanzen in die Natur, sind die Aspekte *Pflanzen* und *Umwelt* angegeben, und zwar in alphabetischer Reihenfolge; über die gegenseitigen Beziehungen der Aspekte wird nichts gesagt.

Die Angabe des Publikationsmediums basiert auf bibliographischen Standards und bildet in der Praxis eine vereinfachte Ausgabe von "Kategoriseringsregler og bibliografisk standard for danske biblioteker"; wir geben die folgenden Publikationsmedien an:

Buch:
- Beitrag im Sammelband
- Hand-/Lehrbuch
- Monographie
- Nachschlagewerk

Periodicum:
- Anzeige
- Artikel
- Leserbrief
- Bericht

Andere Publikationsmedien:
Brief
Laborbericht/-vorschriften
Gesetz/Gesetzesvorlagen/Material von Behörden
Patent
PR-Material

Diese Informationen sind neben Angaben von Autor, Druckjahr, Druckort, Titel, Kommunikationsrichtung u.dergl.m. im bibliographischen Record gesammelt, der mit Hilfe des IBM-Datenverwaltungsprogramms "Arkiv Assistent" gespeichert ist und auf "ISO 690 Documentation — Bibliographic references — Content, form and structure" basiert. (Näheres s. Anhang 2).

5. Der Status eines Korpus im Verhältnis zum Begriff Sprache ist bei jeder Korpusarbeit — sei es bei der Korpuserstellung, sei es bei der Korpusanalyse — eine sehr wesentliche Frage, die aber allzuoft entweder völlig vernachlässigt oder allenfalls nur stiefmütterlich behandelt wird. Nach unserer Auffassung ist das Statusproblem letzten Endes nichts als eine Glaubenssache, und hier folgt somit das Credo des Korpusteams:

Wir nehmen nicht an, daß der Begriff Sprache mit irgend einer Menge von Äußerungen identisch ist, können andererseits auch nicht die Vorstellung akzeptieren, daß sprachliche Äußerungen bloß ein Beweis sind für die Fähigkeit des Menschen zum Sprechen — eine Auffassung, die bedeutet, daß die Struktur des Korpus tatsächlich gleichgültig ist, daß ein Korpus nur als Problemgenerator bei der Prüfung von verschiedenen oft aufgrund konstruierter Sätze aufgestellten Hypothesen verwendbar ist. Wir sind demgegenüber der Ansicht, daß man das Sprachsystem nur durch die Analyse einer größeren Menge von Äußerungen erfassen kann, und müssen deshalb dazu Stellung nehmen, wie sich das Korpus zur umgebenden Sprache verhält, d.h. wir müssen die Repräsentativitätsproblematik berücksichtigen.

Bekanntlich kann man im streng statistischen Sinne des Wortes nur von "Repräsentativität" einer Teilmenge sprechen, wenn sich die Grundmenge definieren läßt. Ein Textkorpus ist selbstverständlich eine Teilmenge, und zwar eine Teilmenge der Grundmenge "alle Äußerungen der jeweiligen Sprache", und diese ist natürlich keineswegs erfaßbar. Betrachten wir Sprache im allgemeinen, erlauben schon ganz banale Faktoren keine Abgrenzung der Grundmenge Sprache, wenn diese diachronisch fundiert ist: teils ist die schriftliche Überlieferung bekanntlich alles andere als vollständig, teils sind mündliche Äußerungen überhaupt nicht festgehalten worden. Begrenzen wir vor diesem Hintergrund die Grund-

menge auf *eine bestimmte Sprache zu einem bestimmten Zeitpunkt*, und unterscheiden wir zwischen *geschriebener* und *gesprochener Sprache*, kommen wir aber nicht weiter. Es ist natürlich unmöglich, alle mündlichen Äußerungen einer gegebenen Periode zu registrieren, und ebenfalls sind nur Teile der geschriebenen Sprache zugänglich; so sind im Gegensatz zu Zeitungen, Magazinen etc. nicht *alle* Broschüren, Pamphlete, Flugblätter, ganz zu schweigen von *allen* Texten privaten Charakters, wie Briefen, Notizen usf. aufspürbar.

In der Beschreibung der vertragsrechtlichen Korpora wird behauptet, man habe einen hohen Grad an Repräsentativität erreicht, und dies läßt sich im Prinzip auch nicht bestreiten; nur ist die Grundmenge, in diesem Fall eine Bibliographie der einschlägigen Rechtsliteratur, durchaus asprachlich, und das Korpus spiegelt damit nicht die Rechtssprache im weitesten Sinne des Wortes wider.

Selbst wenn wir einmal von diesen Problemen absehen, müssen wir mit Rieger (1979) vom Prädikat Repräsentativität Abstand nehmen, denn jede Form von Widerspiegelung der Bibliographie im Korpus beruht auf einer subjektiven Wahl der Eigenschaften, die mit gleicher Häufigkeit in den Texten der Bibliographie und in denen des Korpus auftreten müssen. Wenn man im vertragsrechtlichen Korpusprojekt ein thematisches, in der Praxis ein vertragsrechtlich orientiertes Auswahlkriterium verwendet, wählt man nur eines von vielen möglichen Kriterien. Man hätte sich nämlich ebenfalls eine Selektion vorstellen können nach beispielsweise Dokumententyp, Publikationsmedium, Textsorte oder dem Absender/Empfänger-Verhältnis; das Korpus würde dann einen ganz anderen Inhalt erhalten haben.

Hinzu kommt, daß die Kategorisierung von etwa Dokumententypen und Textsorten natürlich nicht universell, sondern theorieabhängig ist, d.h. durch die Auffassung der Person oder Personengruppe bedingt, die das Korpus erstellt hat. Mit anderen Worten, der Benutzer mag gezwungen sein, sich einem Paradigma zu unterwerfen, das dem eigenen nicht entspricht. Eben aus diesem Grund nahmen wir ebenfalls davon Abstand, die Korpora zu kodieren ("tag"), d.h. die einzelnen sprachlichen Einheiten mit besonderen Kodes für Wortklasse, Gliedfunktion, Satzart etc. zu versehen.

Wir meinen also insgesamt, daß kein Textkorpus das Prädikat *repräsentativ* beanspruchen kann, und die Ergebnisse der Korpusanalysen haben letztendlich auch nur Gültigkeit für die Sprache, wie sich diese im jeweiligen Korpus manifestiert — oder anders ausgedrückt: Das Korpus

ist mit dem Sprachsystem selbst nicht identisch; die allzu wenigen, generell disparaten und zufällig exzerpierten Textbelege früherer Zeiten sind bloß durch große Textmengen ersetzt. Dies bedeutet keine Abkehr von der Arbeit mit Korpora. Wir möchten nur nachdrücklich darauf hinweisen, daß die empirische Korpusanalyse selbstverständlich nicht die nackte Wahrheit enthüllt, sondern weitere theoretische Überlegungen aufgrund eines umfassenden Materials ermöglicht — und dies gilt für alle Bereiche der Linguistik.

Obwohl wir also den Repräsentativitätsgedanken verwerfen müssen, behaupten wir damit nicht, daß unsere Korpora und die unserer Kollegen keine Aussagekraft haben. Schon die intensiven Literaturrecherchen und in unserem Falle der breite Themenzugang lassen vermuten, daß die Textsammlungen eine zuverlässige Widerspiegelung der Sach-/Fachsprache ergeben.

6. Mit dem Voranstehenden haben wir zeigen wollen, daß wir im gentechnischen Korpusprojekt pragmatisch verfahren sind. Nicht aus Not oder Theoriefeindlichkeit; unsere eigenen Erfahrungen sowie unsere Auseinandersetzungen mit den Mitarbeitern anderer Projekte haben uns davon überzeugt, daß die Erstellung eines Korpus eine vor allem praktische Arbeit ist und letztendlich aus einer großen Anzahl von Wahlmöglichkeiten und praktischen Lösungen besteht.

7. Die Korpora stehen allen Sprachforschern kostenlos zur Verfügung. Bestellungen sind an *Ole Lauridsen, Handelshøjskolen i Århus, Fuglesangs Allé 4, DK-8210 Aarhus V* (Tel. +45 86 155588/Fax +45 86 157727) zu richten.

Anhang 1 — Verteilung der Wortstellen und Textexzerpte

COD		BIOTEK.DA	BIOTEK.DE
I	Wortstellen absolut	183.270	217.334
	do. prozentual	18%	22%
	Anzahl Exzerpte	65	87
		Einschl. I(IIa):	Einschl. I(IIa):
		18.430 Wortstellen	53.201 Wortstellen
		8 Exzerpte	24 Exzerpte
			&
			einschl. I(IIb):
			5.288 Wortstellen
			1 Exzerpt
IIa	Wortstellen absolut	376.585	325.026
	do. prozentual	38%	32%
	Anzahl Exzerpte	189	114
		Einschl. IIa(III):	Einschl. IIa(I):
		27.095 Wortstellen	11.751 Wortstellen
		19 Exzerpte	5 Exzerpte
			&
			einschl. IIa(III):
			8.344 Wortstellen
			4 Exzerpte
IIb	Wortstellen absolut	87.160	66.276
	do. prozentual	9%	7%
	Anzahl Exzerpte	15	7
		Einschl. IIb(III):	Einschl. IIb(III):
		12.800 Wortstellen	2.091 Wortstellen
		3 Exzerpte	1 Exzerpt
III	Wortstellen absolut	353.757	393.764
	do. prozentual	35%	39%
	Anzahl Exzperte	254	208
		Einschl. III(IIa):	Einschl. III(IIa):
		26.015 Wortstellen	5.571 Wortstellen
		22 Exzerpte	2 Exzerpte
		&	
		einschl. III(IIb):	
		4900 Wortstellen	
		1 Exzerpt	
	INSGESAMT	1.000.772 Wortstellen	1.002.400 Wortstellen
		523 Exzerpte	416 Exzerpte

Anhang 2 — Bibliographischer Record

PRJ: Project
= Projektname; BIOTEK.DA (dänisch), BIOTEK.DE (deutsch)
NUM: Number
= Identifikationsnummer; 0001-0523: dänisch & 2001- 2416: deutsch
DAT: Date
= Eingabedatum
OPR: Operator
= Operator
PRS: Primary responibility
= Autor
SRS: Subordinate responsibility
= Ergänzende Angaben über Herausgeber, Bearbeiter usw.
TPU: Type of publication
= Publikationsmedium
HST: Host
= Wirt
SBI: Supplementary bibliographic information
= Ergänzende bibliographische Angaben (Druckort, Verlag usw.)
PUY: Year of publication
= Druckjahr
ISN: = ISBN/ISNN-Nummer
EXT: Extent
= Wortstellenanzahl des Gesamtwerks
COR: Corpus
= Wortstellenanzahl des Textausschnitts u. evtl. Seitenangabe
LBR: Library
= Bibliothek
TRL: Translation/Translator
= Übersetzung/Übersetzer
COD: Direction of communication
= Kommunikationsrichtung
ASP: Aspect
= Aspekt
NTS: Notes
= Ergänzende Informationen (z.B. zur Doppelkategorisierung)

Weitere Informationen zu den Korpora:

– Kommen Anmerkungen in einem Text vor, sind diese *im* Text zwischen { } plaziert
– Die Anzahl und Art von Bildern, Graphen etc. ist mit Ziffern und Sonderzeichen zwischen Identifikationsnummer und ASCII-Text angegeben; die Zeichen sind:

@ = Bild mit Text
# = Bild ohne Text
¤ = Abbildung ohne Text
§ = Abbildung mit Text
\ = tabellarische Übersicht
¢ = längere Formel

Nur in den Fällen, wo Angaben zur Plazierung einer Illustration als wesentlich für das Verständnis eines Textes betrachtet wurde, ist das Zeichen *im* Text angebracht worden; auf Angaben zwischen Identifikationsnummer und Text wurde dabei nicht verzichtet.

– Griechische Buchstaben kommen in vielen Formeln vor. Ihre Namen sind in den Texten ausgeschrieben, und vor dem Namen steht dann das Zeichen ¢. ¢ hat damit zwei Funktionen.

Literatur (Auswahl):

Bausch, Karl-Heinz (1975) "Zur Problematik der empirischen Basis in der Linguistik. Diskutiert am Modusgebrauch in Konditionalsätzen". In: *Zeitschrift für Germanische Linguistik* 3, 123-148

Bergenholtz, Henning (1988) "DK 88: Et korpus for dansk almensprog". In: *HERMES 1*, 229-237.

Bergenholtz, Henning (1989) "Korpusproblematik in der Computerlinguistik. Konstruktionsprinzipien und Repräsentativität". In: Batori, Istvan et al. (eds.) *Computational Linguistics. Ein internationales Handbuch computerunterstützter Sprachforschung und ihrer Anvendung*. Berlin/New York: de Gruyter

Dyrberg, Gunhild, Dorrit Faber, Steffen Leo Hansen og Joan Tournay (1988) "Etablering af et juridisk tekstkorpus". In: *HERMES 1*, 209-227

Hahn, Walther v. (1983) *Fachkommunikation. Entwicklung — Linguistische Konzepte — Betriebliche Beispiele*. Berlin/New York (= Sammlung Göschen 2223)

Lauridsen, Karen M. og Ole Lauridsen (1989) "Tekstkorpora. En ny forskningsaktivitet ved Handelshøjskolen". In: *Festskrift i anledning af Handelshøjskolens 50-års jubilæum 31. august 1989*. Handelshøjskolen i Århus

Rieger, Burghard (1979) "Repräsentativität. Von der Unangemessenheit eines Begriffes zur Kennzeichnung eines Problems linguistischer Korpusbildung". In: Bergenholtz, Henning og Burkhard Schaeder (eds.) *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora*. Königstein/Ts.