Korpora als Ausgangspunkt für die Extraktion von terminologischen Daten

Heribert Picht

Norges Handelshøyskole

1. Einleitung

Ein Korpus kann ganz allgemein als eine Sammlung von Dokumenten definiert werden, die nach bestimmten Auswahlkriterien so zusammengestellt ist, daß sie eine Gesamtheit bildet und für einen oder mehrere Zwecke verwendbar ist. Eine solche Definition sagt nichts über die Form an sich aus und beinhaltet auch nicht, daß ein Korpus in maschinenlesbarer Form vorliegen muß. Im Grunde haben so gut wie alle Untersuchungen auch vor dem Computerzeitalter sich auf Textsammlungen – Korpora – der einen oder anderen Art gegründet. Mit anderen Worten: der Korpusgedanke an sich ist nicht neu.

Daß maschinenlesbare Korpora wesentliche Vorteile bieten ist unumstritten. Die Bearbeitung von großen Datenmengen, die eine breitere Grundlage der Untersuchungen erlauben, die schnelle Bearbeitung und die Anwendung für mehrere Zwecke sind allgemein anerkannte Vorteile. Die ersten beiden Punkte sind quantitativer Art, die einen direkten oder indirekten Einfluß auf die Qualität der Untersuchung haben können. Man sollte jedoch nicht vergessen, daß sie in erster Linie aber doch nur rein quantitativer Art sind, da Computerprogramme, die zur Bearbeitung von Korpora herangezogen werden können (1) keine analytisch-kognitiven Eigenschaften haben und (2) daß ein Korpus nicht mehr liefern kann, als es enthält. Die letztgenannte Beschränkung kann bis zu einem gewissen Grade durch die Menge der Daten ausgeglichen werden, aber auch hier sind Grenzen zu erwarten.

In diesem Beitrag soll nicht in die Diskussion eingetreten werden, inwieweit die Korpuslinguistik eine selbständige Disziplin ist oder nicht. Siehe hierzu Hansen (1988).

Hansen (1988) hat eine Reihe von Definitionen des Begriffs 'Korpus' untersucht und konnte folgende gemeinsame Merkmale feststellen:

- eine endliche Menge von sprachlichen Daten in Form von gesprochenen oder geschriebenen Texten oder Textteilen;
- ein für einen bestimmten Zweck zusammengestellter Ausschnitt aus einem Gegenstandsbereich;
- eine Teilmenge von Daten aus einer Grundmenge, so daß alle Elemente der Grundmenge mit gleicher statistischer Wahrscheinlichkeit in der Teilmenge enthalten sind und daß diese Elemente sich gleichmäßig auf die gesamte Grundmenge verteilen;
- eine im Verhältnis zu einer gegebenen Grundmenge repräsentative Teilmenge.

Hieraus leitet er fünf generelle kategorielle Eigenschaften eines Korpus ab:

1. Funktion, wobei er zwischen der Anwendungsfunktion und der methodischen Funktion unterscheidet;
2. Inhalt
3. Umfang
4. Struktur
5. Repräsentativität.

Die ersten vier Eigenschaften sind einleuchtend und sollen hier nicht näher behandelt werden.

Der Frage der Repräsentativität muß jedoch besonders im Hinblick auf eine terminologische Nutzung von Korpora besondere Beachtung geschenkt werden. Hansen unterscheidet zwischen:

1. Statistischer Repräsentativität, die zu statistisch repräsentativen Korpora führt;
2. Qualitativer Repräsentativität, die qualitativ repräsentative Korpora ergibt;
3. Themendeckende Repräsentativität, die themendeckende Korpora erfordert.

Alle drei Formen der Repräsentativität sind relevant für die korpusbasierte Terminologiearbeit.

Je nach Zweck und Untersuchungsziel sind auch Einteilungen nach anderen Kriterien relevant, z.B. wird unterschieden zwischen:

1. Polyfunktionalen und monofunktionalen Korpora
2. Teiltextkorpora (sample corpus) und Gesamttextkorpora (monitor corpus)
3. Gemeinsprachlichen und fachsprachlichen Korpora.

Nach diesen sehr generellen Bemerkungen werde ich mich eher praktisch-terminologischen Fragestellungen zuwenden und folgende Teilthemen herausgreifen:

1. Terminologisch relevante Datentypen und ihre Repräsentationsformen;
2. Korpora für bestimmte Zwecke
3. Alter und Aktualität von Texten
4. Fachliche Kompetenz der Textverfasser - Texttypen
5. Korpora für mehrsprachige Terminologiearbeit
6. Wissensstrukturen in Korpora und terminographischen Produkten.

2. Korpora und Terminologiearbeit

2.1 Terminologisch relevante Datentypen und ihre Repräsentationsformen

Ein Korpus besteht aus einer endlichen Menge expliziter und impliziter Repräsentationsformen, die für die Terminologiearbeit von Bedeutung sein können – das ist banal. Aber welche Repräsentationsformen für terminologische Daten kommen in Betracht ?

Die folgende Übersicht über terminologisch relevante Daten läßt sich aus der Norm 'ISO 12 620 – Computer applications in terminology – Data categories' ableiten.

Repräsentationsformen

	Sprachliche	Nicht-prachliche
Gegenstand	Namen Faktenbeschreibungen	Bilder
Begriff	Benennungen und Zusatzinformationen z.B. gram. Art Formeln Paraphrasen Phraseologische Einheiten	Formeln Symbole graphische Repräsentationen
	Definitionen Erklärungen	
Beziehungen	verbal ausgedrückt	graphisch dargestellt z.B. Begriffssysteme

Normalerweise enthalten Korpora nur Texte. Korpora, die Texte und Abbildungen enthalten, sind noch die Ausnahmen und ihre Nutzung befindet sich im Versuchsstadium.

Welche terminologischen Daten können über ihre Repräsentationsformen aus einem Korpus gewonnen werden?

1. Einwortbenennungen können problemlos durch Zeichenvergleich festgestellt werden, allerdings muß intellektuell sortiert werden, ob es sich um Benennungen im terminologischen Sinne handelt.
2. Mehrwortbenennungen lassen sich durch Konkordanzen ermitteln, allerdings
   ist es nicht selten erforderlich, die Grenzen der Benennung intellektuell zu bestimmen.

3. Phraseologische Einheiten können durch die Suche der 'Kernbenennung' und ihrer Umgebung festgestellt werden; auch hier ist die Interaktion zwischen Mensch und Maschine unvermeidlich.
4. Die Paraphrase eines Begriffs kann nur durch Indikatoren wie 'kann ... umschrieben werden' festgestellt werden, da ein direkter Zugriff nicht möglich ist.
5. Definitionen und Erklärungen können ebenfalls nur durch Indikatoren, zumeist Verben wie 'definieren, bedeuten, erklären' ermittelt werden. Auch hier ist ein unmittelbarer Zugriff ausgeschlossen.
6. Beispiele fallen ebenfalls in die beiden letztgenannten Kategorien.
7. Graphische Darstellungen und Bilder können, wenn überhaupt, nur durch Indikatoren wie 'Fig. x' festgestellt werden und sind dann oft auch nur im Originaldokument zugänglich.

Zusammenfassend kann festgestellt werden, daß es zwei Arten der Datenextraktion gibt: 1. durch direkten Zeichenvergleich und 2. durch Interaktion verschiedener Art mit dem Computer. Zwar gibt es heute schon eine Reihe von Programmen, die diese Interaktion erleichtern, aber sie können sie nicht vollständig ersetzen. Zu den derzeitigen und eventuell zukünftigen Möglichkeiten, die genannten Datentypen zu extrahieren siehe Estopà (2000). In allen Fällen werden die elektronischen Werkzeuge jedoch immer nur terminologische Rohdaten liefern können.

2.2 Korpora für bestimmte Zwecke

Folgt man Hansens obengenannter Einteilung und setzt sie in Beziehung zur terminologischen Nutzung von Korpora, ergibt sich folgende Bild:

1. **Statistisch repräsentative Korpora** haben sich besonders für die Erarbeitung von didaktischem Material bewährt. Hier wären u.a. Hoffmanns Fachwortminima für bestimmte Fachgebiete z.B. 'Fachwortschatz Bauwesen. Häufigkeitswörterbuch. Russisch, Englisch, Französisch.' zu nennen, deren Zweck es ist, dem Benutzer die hochfrequenten Benennungen und ihre Äquivalente zugänglich zu machen. Der Deckungsgrad dieser Minima liegt bei etwa 85 %.

2. **Qualitativ repräsentative Korpora**. Ein wichtiges Kriterium für jegliche terminologische Arbeit ist es, daß das terminographische Endprodukt so

vollständig wie möglich ist. Das bedeutet, daß das Korpus in der Regel sehr umfangreich sein muß, da neben der rein thematischen Abdeckung auch eine texttypologische Deckung gegeben sein sollte.Texttypologische Deckung umfaßt sowohl die fachliche Qualität der Texte als auch die pragmatischen Dimensionen, die z.B. durch verschiedene Fachlichkeitsgrade (Hoffmann 1984:65) und die entsprechenden Kommunikationssituationen zum Ausdruck kommt. Ferner muß den terminologischen Abweichungen in großen Sprachgemeinschaften, z.B. Spanisch in Spanien und in verschiedenen südamerikanischen Ländern, Rechnung getragen werden.

3. **Themendeckende Korpora.** Einige Verfasser verwenden die Bezeichnung 'themenrelatierte Korpora' als Synonym. Diese Verwendung erscheint mir unzweckmäßig, da gerade im terminologischen Bereich die Vollständigkeit, die fachliche Deckung eines definierten Fachbereiches, ein anzustrebendes Ziel ist. 'Themenrelatiert' dagegen ist vage, schwächt die Forderung nach Vollständigkeit stark ab und besagt eigentlich nur, daß ein Korpus eine thematische Beziehung zu einem bestimmten Thema (Fachbereich) hat.

   In diesem Zusammenhang stellt sich auch die Frage nach der Anwendbarkeit von Teiltextkorpora für die terminologische Arbeit. Durch die Auswahl von Teiltexten geht in der Regel eine nicht unerhebliche Menge von terminologisch relevanter Information verloren. Solche Korpora sind zwar 'themenrelatiert' aber nicht 'themendeckend' und kommen damit für qualitativ hochstehende Terminologiearbeit kaum in Frage. Daraus ergibt sich, daß für die terminologische Arbeit nur ein Gesamttextkorpus geeignet sein kann. Gesamttextkorpora können sehr verschiedener Art sein. Nachstehend seien einige Beispiele genannt.

1. <u>Quantitativ und qualitativ themendeckende Korpora</u>

   A. Ein Korpus, das aus dem Gesamtwerk eines Verfassers besteht, z.B. Niels Bohrs gesamtes Schrifttum zur Atomphysik. Ein solches Korpus ist z.B. geeignet, diachronische Untersuchungen zur Begriffsentwicklung durchzuführen. (Ahmad, Jensen 1998:20ff).

   B. Ein Gesamttextkorpus bestehend aus:

1. **einer** Texttype, die **ein** Fachgebiet und **eine** Sprachregion abdeckt. Man könnte z.B. ein Korpus der DIN-Normen zum Fachgebiet 'Abwassertechnik' zusammenstellen;

2. **einer** Texttype, die **ein** Fachgebiet und **mehrere** Sprachregionen einer Sprachgemeinschaft abdeckt. Ein solches Korpus könnte z.B. aus allen nationalen spanischsprachigen Normen der Serie ISO 9000 bestehen oder aus den gesamten Gesetzestexten der selben Sprachregion zum Thema 'Menschenrechte';

3. **mehreren** Texttypen, die **ein** Fachgebiet und **mehrere** Sprachregionen abdecken. Korpora dieser Art könnten z.B. für das gesamte Schrifttum einer philosophischen Schule oder für die gesamte Dokumentation eines Unternehmens zu einem bestimmten Produkt zusammengestellt werden.

2. Statistisch und qualitativ themendeckende Korpora

Ein Gesamttextkorpus bestehend aus:

A. **mehreren** Texttypen, die **ein** Fachgebiet und **eine** Sprachregion abdecken, würde mehrere verschiedene Repräsentationsformen zu einem Begriff anbieten und wäre besonders für die deskriptive Terminologiearbeit geeignet, die gleichzeitig eine größere Bandbreite der pragmatischen Erfordernisse abdecken könnte.

B. **mehreren** Texttypen, die **ein** Fachgebiet und **mehrere** Sprachregionen abdecken, würde die Anzahl der Repräsentationsformen für einen Begriff weiter erhöhen und zu einer qualitativ besseren kontrastiv-deskriptiven intralingualen Terminologiearbeit wesentlich beitragen.

2.3 Alter und Aktualität von Texten

Alter im Verhältnis zur Aktualität von Texten ist ein unklarer Begriff. Es ist daher besser vom Wissensstand, den ein Text repräsentiert, zu sprechen. Eine Altersangabe allein ist kein verläßliches Kriterium für die Aufnahme oder den Ausschluß eines Textes. Entscheidend ist vielmehr, wie rasch ein Fachgebiet sich entwickelt und wie hoch die Alterungsgeschwindigkeit des Wissensstandes in einem Text ist. Mit anderen Worten: die Entwicklungsgeschwindigkeit eines

Fachgebietes ist ausschlaggebend für die Auswahl der in Korpora aufzunehmenden Texte.

Für die Erstellung eines Korpus bedeutet das, daß Fachleute eine entscheidende Rolle bei der Auswahl der Texte spielen müssen, denn nur sie sind in der Lage, die fachliche Repräsentativität eines Korpus zu sichern.

Ein weiterer wichtiger Punkt ist die Aktualisierung von Korpora, wenn es darum geht, ein Korpus auch für die Aktualisierungsarbeit von terminologischen Beständen zu nutzen. Ein fachlich veraltetes Korpus kann einem veralteten Wörterbuch gleichgestellt werden. Allerdings sieht man in der Praxis nur wenige Fälle von gezielter, fachlicher Korpusaktualisierung.

2.4 Fachliche Kompetenz der Textverfasser - Texttypen

Der gemeinsame Nenner von Fachtexten, die ein Korpus bilden sollen, ist ihre wissensvermittelnde und kommunikative Funktion. Es sind verschiedene, oft sehr einfache Akteurmodelle in der Fachkommunikation vorgelegt worden. Meistens beschränken sie sich auf Kombinationen wie

Fachmann - Fachmann
Fachmann - Laie
Laie - Fachmann
Laie - Laie.

Für die Korpuspraxis ist dieser Raster zu grob und entspricht nicht der Wirklichkeit und damit auch nicht den auszuwählenden Texten. Da Wissenstransfer auch gleichzeitig den Begriff 'Wissensgefälle' bedingt, sind differenziertere Modelle erforderlich (Picht 1999:29ff). Für die Zusammenstellung fachlicher Korpora bedeutet das, daß fachlich unzureichende Texte ausgeschlossen werden müssen, da sie keine Garantie für fachliche Korrektheit bieten. Kaufmann (1992:67) unterstreicht zu Recht: "Wenn ein fachliches Korpus in fachlichem Zusammenhang als Werkzeug dienen soll, muß es auf korrekter Grundlage aufgebaut sein" (meine Übersetzung). Kaufmann konnte eklatante fachliche Fehler in Texttypen nachweisen, die von Laien (Jounalisten) für Laien geschrieben worden waren.

2.5 Korpora für mehrsprachige Terminologiearbeit

In der Regel wird ein Korpus für eine Sprache zusammengestellt. Für die mehrsprachige Terminologiearbeit sind daher Korpora in jeder zu bearbeitenden Sprache erforderlich. Was die inhaltliche Vergleichbarkeit solcher Korpora betrifft, wird oft davon ausgegangen, daß sie bei naturwissenschaftlich-technischen Korpora relativ hoch ist. Diese Annahme darf aber nicht als Axiom gelten, denn je nach Fachgebiet haben sich erhebliche Unterschiede herausgestellt, z.B. Pflugtypen in verschiedenen Ländern oder Dachformen und deren Materialien in verschiedenen geographischen Regionen.

Für die eher gesellschaftswissenschaftlichen Fachgebiete ergeben sich auf jeden Fall größere Schwierigkeiten bei der Erstellung von 'inhaltlich gleichwertigen' Korpora. Der Vergleich der aus ihnen zu gewinnenden terminologischen Daten ist schon wegen der starken Bindung der Daten an gesellschaftliche und nationale Systeme und Strukuren nur intellektuell möglich, da ein eingehendes z.B. rechtsvergleichendes Hintergrundwissen erforderlich ist, das die Texte in der Regel nicht enthalten.

Daraus ergibt sich, daß Korpora zwar monolinguale terminologische Rohdaten liefern können, der terminologische Vergleich und damit die mehrsprachige Arbeit jedoch intellektuell erfolgen muß.

2.6 Wissensstrukturen in Korpora und in terminologischen Produkten

Jeder Text hat seine eigene Wissensstrukturierung in Übereinstimmung mit dem kommunikativen Ziel des Textes. Diese Wissensstrukturen entsprechen fast nie der begriffsbasierten Struktur in terminographischen Produkten, auch wenn einzelne Strukturschritte verbalisiert sein können und in terminographischen Produkten unmittelbar verwendbar sind. Daraus ergibt sich, daß aus einem Korpus nur 'Wissensbausteine' als terminologische Datenelemente oder Teile von ihnen über ihre Repräsentationsformen auffindbar sind und entnommen werden können. Es ergeben sich wie gesagt Rohdaten, die durch die terminologische Analyse und Weiterbearbeitung in eine terminologische Wissenstruktur überführt werden müssen. Diese Struktur muß dann allerdings alle terminologisch relevanten Daten enthalten, die es dem Benutzer eines terminographischen Produktes ermöglichen, sie erneut in einen Text mit bestimmtem Kommunikationsziel korrekt einzusetzen.

3. Konklusion

Es kann festgestellt werden, daß folgende Punkte für die Zusammenstellung von Korpora für terminologische Zwecke zu beachten sind:

1. Beschaffenheit des Korpus
- es sollte möglichst immer ein Ganztextkorpus sein
- ein möglichst hoher Grad an statistischer Repräsentativität ist erforderlich
- das Korpus sollte themendeckend (fachbereichdeckend) sein
- hohe qualitative Repräsentativität ist unerläßlich
- eine laufende Aktualisierung ist unumgänglich, wenn der Wert des Korpus erhalten bleiben soll.

2. Auswahl der Texte
- die Auswahl sollte immer in Zusammenarbeit mit Fachleuten erfolgen
- die wissenmäßige Aktualität ist ausschlaggebend, das Alter dagegen untergeordnet.

3. Werkzeuge zur Korpusnutzung
- die wesentlichste Beschränkung der Werkzeuge liegt z.Z. in ihren fehlenden analytisch-kognitiven Eigenschaften
- die Werkzeuge sind in den letzten Jahren erheblich verbessert worden und erleichtern wesentlich die Extraktion von Daten, die jedoch immer nur terminologische Rohdaten sein können
- die Schnittstelle 'Werkzeug – Mensch' ist analytisch-kognitiv bedingt
- die Schnittstelle ist durch die computerlinguistische Forschung in ständiger Bewegung, doch gibt es m.E. eine grundlegende Grenze, solange den Werkzeugen analytisch-kognitive Eigenschaften fehlen.

4. Wirtschaftliche Anwendung von Korpora
- m.E. besteht noch ein großer Abstand zwischen der Forschung und der wirtschaftlichen Anwendung
- soweit mir bekannt, hat man noch keine Kosten-Nutzen-Analysen durchgeführt, die belegen können, wie groß die Arbeitsersparnis durch den Einsatz von Korpora in der Produktion von qualitativ hochwertigen Terminologien ist oder sein kann. Daraus ergibt sich die fast ketzerische Frage: Ist es z.B. für ein Unternehmen oder eine Institution/Organisation wirtschaftlich von Vorteil, Korpora für die Terminologiearbeit aufzubauen, für eine optimale Nutzung aufzubereiten und zu unterhalten?

Zweifellos wird man mir entgegenhalten, daß Korpora ein wichtiger Schritt zur Erleichterung der terminologischen Arbeit sind. Dem kann ich unbedingt zustimmen, doch sollte man auch die Beschränkungen erkennen, die trotz aller Verbesserungen der Werkzeuge bei der Nutzung von Korpora weiterhin bestehen.

Bibliographie

Ahmad, K.; Jensen, L.(1998): *En lille smule om atomernes bygning: Niels Bohr Writing Atomic Structure*. In: Terminology Science & Research; vol. 9(1998), no. 2. Wien: International Network for Terminology (TermNet). 20 – 33.

Estopà, R.(2000): *Extracción de terminología: elementos para la construcción de un SEACUSE (Sistema de Extracción Automática de Candidatos a Unidades de Significación Especializada)*. Tesis doctoral. Institut Universitari de Lingüística Aplicada. Universidad Pompeu Fabra. Barcelona.

Hansen, Steffen L. (1988): *Korpuslinguistik. Teori – metode – praksis*. LAMBDA nr. 5. Institut for Datalingvistik, Handelshøjskolen i København.

ISO 12 620 Computer applications in terminology – Data categories. First edition 1999.

Kaufmann, U. (1992): *Anvendelse af det danske genteknologiske tekstkorpus ved udarbejdelse af Genteknologisk Ordbog, med specielt henblik på udvælgelsen af eksempler*. In: Proceedings af Seminar om Korpuslingvistik i Fagsprogsforskning. Hindsgavl Slot, 26. og 27. Nov. 1992. Hindsgavl: [s.n.]. 56 – 68.

Picht, H. (1999): *Die Begriffe 'Fachmann' und 'Laie' in der Fachkommunikation*. In: Internationale Wirtschaftsbeziehungen: Mehrsprachige Kommunikation von Fachwissen; W. Wieden, A. Weiss (Hrsg.). Göppingen: Kümmerle Verlag, 29 – 42.

Stummann, B. M. (1992).: *Anvendelsesmuligheder og faglig indhold af det danske genteknologiske tekstkorpus*. In: Proceedings af Seminar om Korpuslingvistik i Fagsprogsforskning. Hindsgavl Slot, 26. og 27. Nov. 1992. Hindsgavl: [s.n.]. 69 – 74.