Andrea Lehr: Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze.

- Abstract -

Ausgangspunkt - Fragestellung - Zielsetzung

Ausgangspunkt der vorliegenden Arbeit ist die lexikorientierte Kollokationstheorie des britischen Kontextualismus, deren Grundstein John Rupert Firth legte. Dort werden Kollokationen als Zusammenfassungen von sprachlichen Einheiten betrachtet, die in konkreten Texten miteinander vorkommen und sich in syntagmatischer Nähe zueinander befinden. Kollokationen im kontextualistischen Sinne haben meist nur wenig mit Kollokationen im Sinne der Phraseologieforschung gemein. Nach ersterer Auffassung sind sie Entitäten, die zu Zwecken sprachwissenschaftlicher Beschreibungen aus den zu analysierenden Texten herausgefiltert werden. Nach zweiterer Auffassung dagegen sind sie im Graubereich zwischen freien syntagmatischen Verbindungen und Phrasemen angesiedelt und zählen eher zu den Einheiten des Sprachsystems.

Unser erstes Anliegen besteht darin zu zeigen, daß sich die Kollokationstheorie des britischen Kontextualismus in hervorragender Weise zur Bewältigung einiger *Aufgabenstellungen der maschinellen Sprachverarbeitung* nutzen läßt. Während die maschinelle Sprachverarbeitung große Erfolge auf dem Gebiet der Syntax verzeichnen kann, gibt es im Bereich der Lexik eine Reihe ungelöster Probleme. So ist die maschinelle Identifizierung von sprachlichen Einheiten dann kaum zu leisten, wenn diese Einheiten mit anderen sprachlichen Einheiten formidentisch sind oder sich aus mehreren durch Blanks getrennten Komponenten zusammensetzen. Die noch immer bevorzugten Strategien zur Lösung dieser Probleme bestehen darin, entweder restriktive Regelungen für die Erstellung von Texten, die einer maschinellen Bearbeitung zugedacht sind, zu entwickeln oder das enzyklopädische Wissen, über das Menschen verfügen, in formalen Systemen zu kodifizieren und dann mit sprachlichen Einheiten zu verweben. Die eine Lösungsstrategie halten wir für schlicht inhuman, die andere erscheint uns ineffektiv, da in einer formalen Repräsentation enzyklopädisches Wissen wiederum nur aus sprachlichen Ausdrücken oder den Symbolen einer eigens zu diesem Zweck geschaffenen Beschreibungssprache bestehen kann. Trotz dieser "Versprachlichungszwänge" können die auf sprachexternen Objekten und Objektkonstellationen basierenden Beschreibungen einer solchen Repräsentation nur durch den Aufbau äußerst komplexer Regelsysteme den entsprechenden sprachlichen Einheiten zugeordnet werden.

Entgegen der beiden genannten Vorgehensweisen wollen wir sprachliche Einheiten nicht durch die Beschreibung sprachexterner Gegebenheiten charakterisiert wissen, sondern anhand ihres textuellen Verhaltens, wie es sich mit Hilfe rein formlicher Kriterien wie "graphische Gestalt" und "Textposition" ermitteln läßt. Hier sind wir wieder bei der kontextualistischen Kollokationstheorie angelangt, denn unter textuellem Verhalten ist nichts anderes zu verstehen als: "In welchen Kollokationen kommt welche Einheit mit welcher Häufigkeit und mit welcher Signifikanz vor?"

Unser zweites Anliegen besteht darin, ein operationales Sprachanalysemodell (vgl. Fußnote 1) auf der Ebene der Lexik zu entwickeln, welches zeigt, wie - von umfangreichen Textkorpora ausgehend - Beschreibungen des textuellen Verhaltens sprachlicher Einheiten maschinell erzeugt werden können und wie diese in späteren Anwendungen

zu nutzen sind; Anregungen dazu gaben bereits erprobte Analyseverfahren von Godelieve Berry-Rogghe, Peggy Haskel, Göran Kjellmer, Kathleen R. McKeown und Frank Smadja sowie von John McH. Sinclair.

Eine Besonderheit unseres Modells ist es, auf zwei unterschiedlichen Originalen zu fußen. Als präexistentes Original zu diesem Modell fungiert die kontextualistische Kollokationstheorie, die modellintern in spezifischer Weise beschrieben wird. Als projektiertes, erst noch zu entwickelndes Original fungieren dagegen konkrete maschinelle Umsetzungen des KOLYNET-Modells. Von daher ist das Modell (ebenso wie die gesamte Arbeit) im Grenzbereich zwischen Sprachwissenschaft und Computerlinguistik anzusiedeln und soll die Theorien der einen wissenschaftlichen Disziplin für die Aufgaben der anderen nutzbar machen.

Untersuchungsgang

Die Arbeit gliedert sich in zwei Großkapitel und ein Zwischenkapitel:

- Das erste Großkapitel (Kapitel 2: *Kollokationen - Theorien, Methoden, Termini*) bietet einen wissenschafthistorischen Überblick über das Gebiet der Kollokationen.
- Das zweite Großkapitel (Kapitel 4: *Das KOLYNET-Analysemodell*) befaßt sich mit der Entwicklung eines operationalen Kollokationsanalysemodells, welches letztlich der maschinellen Erzeugung lexikalischer Netze dienen soll.
- Das Zwischenkapitel (Kapitel 3: *Präliminarien zu KOLYNET*) schließlich erörtert diejenigen Sachverhalte, die vor der Konzeptualisierung von KOLYNET einer Klärung bedürfen, aber nicht direkt zur kollokations- und sprachtheoretischen Basis von KOLYNET gehören; darunter fallen u.a. Fragen der Korpusgestaltung und Korpusaufbereitung sowie Fragen der Organisation relationaler Datenmodelle.

Im ersten Großkapitel befassen wir uns mit unterschiedlichen Kollokationstheorien und der Theorie der lexikalischen Selektionsrestriktionen, die oft mit ersteren gleichgesetzt wird. Im Blickpunkt unseres Interesses stehen dabei die Kollokationstheorie des britischen Kontextualismus und - da diese sprachwissenschaftliche Schule außerhalb Großbritanniens wenig bekannt ist - die grundlegenden theoretischen und methodologischen Implikationen kontextualistischer Sprachforschung. Generell zeichnet sich diese Schule durch das Postulat eines strikt empirischen Vorgehens aus, was dem skizzierten Vorhaben natürlich entgegenkommt und einen wesentlichen Grund für unsere Konzentration auf die kontextualistische Kollokationstheorie darstellt.

Aufgrund ihres Empiriepostulats entwickelte die kontextualistische Schule ein ausgefeiltes methodologisches Instrumentarium zur manuellen Analyse konkreter Texte. Den Teil dieses Instrumentariums, der sich auf Kollokationsanalysen bezieht, wollen wir nun kurz vorstellen (wobei wir der Einfachheit halber anstelle der englischsprachigen Termini die von uns verwendeten und in manchen Fällen eigens gebildeten deutschsprachigen Termini anführen):

- Aus kontextualistischer Sicht besteht ein Text auf der lexikalischen Betrachtungsebene aus einer syntagmatischen Reihung von *Wörtern*.

- Gruppen dieser Wörter lassen sich innerhalb eines Textes zu *Kollokationen* zusammenfassen. Dabei gilt, daß Kollokationen intern in zwei hinsichtlich ihres Status unterschiedliche Teile zerfallen.
- Der *Kollokant* ist das Wort einer Kollokation, von dem die Betrachtung ausgeht.
- Das Wort oder die Wörter, die im Kollokatsbereich zu diesem Kollokanten liegen, bilden sein *Kollokat*. Dieses dient der Feststellung singulärer Kookkurrenzphänomene.
- Auf der paradigmatischen Ebene können diejenigen Kollokate, deren Kollokanten als identisch gelten, zu einem *Kollokatspotential* zusammengefaßt werden. Ein Kollokatspotential dient zur Charakterisierung von identischen Kollokanten.
- Identische Kollokanten können unter einem virtuellen Repräsentanten, dem *Kollokantem*, zusamengefaßt werden.
- Kollokanteme wiederum können anhand der ihnen zugeordneten Kollokatspotentiale miteinander verglichen und auf ihre Ähnlichkeit hin untersucht werden. Als ähnlich geltende Kollokanteme werden in *Kollokantemgruppe* zusammengefaßt.

Nach Abschluß einer solchen Kollokationsanalyse findet sich jedes Wort, das in den Status eines Kollokanten gelangte, durch die Verknüpfung mit einem Kollokatspotential sowie die indirekte Zugehörigkeit zu einer Kollokatemgruppe lexikalisch charakterisiert.

In Anlehnung an das dargelegte kontextualistische Instrumentarium für manuelle lexikalische Analysen entwickeln wir das KOLYNET-Analysemodell. Es ist dazu gedacht, eine konzeptionelle Basis für maschinelle Analysen über großen maschinenlesbaren Korpora zu bieten und besteht aus einer Reihe unterschiedlicher Subverfahrensmodelle. Diese zeigen, wie nach und nach aus den verwendeten Korpora ein lexikalisches Netz (vgl. Fußnote 2) erstellt werden kann.

Auf den ersten Blick scheint die Kollokationsanalysemethode der kontextualistischen Schule ohne nennenswerte Schwierigkeiten in ein operationales Modell überführbar zu sein, beim genaueren Hinsehen ergeben sich jedoch einige Schwierigkeiten, für die es im zweiten der genannten Großkapitel Lösungen zu entwickeln gilt.

Die erste Schwierigkeit besteht darin, daß die kontextualistische Schule trotz der Bezugnahme auf Texte Wörter als Ausgangseinheiten der Analyse betrachtet. Wörter sind jedoch Einheiten des Sprachsystems, nicht Einheiten konkreter Äußerungen. Zudem macht eine Buchstabenfolge allein nicht in jedem Falle deutlich, die Ausprägung welchen Wortes sie darstellt. Zu einer solchen Feststellung bedarf es häufig einer Einbeziehung des sprachlichen Kotextes und/oder des situativen Kontextes. Wir gehen deshalb von uninterpretierten Buchstabenfolgen aus (in Anlehnung an die terminologischen Vorgaben der kontextualistischen Schule *Strukturwortformen* genannt), von denen wir zunächst einmal nichts wissen, außer daß sie (aller Wahrscheinlichkeit nach) Realisierungen von Wörtern sind.

Strukturwortformen sind in unserem Modell basale Analyseeinheiten der Abstraktionsstufe 1 und können durch Zusammenfassungen aufgrund von Formidentität in Analyseeinheiten der Abstraktionsstufe 2, in sogenannte Wortformen, überführt werden. Schließlich haben wir noch eine Abstraktionsstufe 3 vorgesehen, auf der Wortformen zu

Hypersegmenten zusammengefaßt sind. Hypersegmente unterscheiden sich von Wörtern dadurch, daß ein Hypersegment alle mit ihm formidentischen Wörter in sich vereint (vgl. Fußnote 3). Auch Wörter, die nicht formidentisch miteinander sind, werden zu einem Hypersegment zusammengefaßt, wenn die ihnen zuzuordnenden Wortformenklassen formidentische Elemente aufweisen. Ein Vorschlag für ein Wortformeninventar, welches die Zuordnung von Wortformen zu Hypersegmenten leistet, findet sich in Kapitel 3.

Ein zweite Schwierigkeit ergibt sich dadurch, daß nach kontextualistischer Auffassung Phraseme etc. als je eine sprachliche Einheit zu betrachten sind. Auch hier gestaltet sich die maschinelle Identifizierung sehr schwierig. Da sie erst im Zuge einer Kollokationsanalyse geleistet werden und deshalb nicht deren Voraussetzung sein kann, haben wir zu Beginn von KOLYNET auf eine gesonderte Behandlung von Phrasemen und anderen Mehrwortlexemen verzichtet.

Die dritte Schwierigkeit besteht in der willkürlichen Festsetzung von Kollokatsbereichsgrenzen. Wir haben uns dafür entschieden, von einem relativ kleinen Kollokatsbereich mit nur sechs Textpositionen auszugehen und zu einem späteren Zeitpunkt der Analyse gezielt nach Kollokationen innerhalb eines erheblich erweiterten Kollokatsbereichs zu suchen.

Die vierte Schwierigkeit schließlich ist, den Ähnlichkeitsgrad zwischen Kollokatspotentialen und damit auch zwischen Kollokantemen zu bestimmen. Um dies zu leisten, haben wir zuerst eine Formel zur paarweisen Ähnlichkeitsberechnung entwickelt und anschließend anhand spezifischer Regeln einander in besonderem Maße ähnliche Paare von Kollokantemen in Kollokantemgruppen überführt.

Ergebnisse

Die mit KOLYNET projektierten Tabellen repräsentieren in erster Linie die Analyseeinheiten der drei genannten Abstraktionsstufen:

- als singuläre *Strukturwortformen*, *Wortformen* und *Hypersegmente* sowie gleichzeitig als *Kollokanteme*,
- als *Kollokate*, *Kollokanten* und zusammengefaßt zu *Kollokationen*,
- als Elemente von *Kollokatspotentialen* und *Kollokantemgruppen*.

Dabei schlagen wir vor, zunächst alle Strukturwortformen der zu analysierenden Korpora nacheinander zu Kollokanten zu deklarieren und alle dadurch ermöglichten Kollokationen innerhalb des zulässigen Kollokatsbereichs zu bilden. Anschließend sind mit speziellen statistischen Formeln die Signifikanzwerte zu den gebildeten Kollokationen zu errechnen. Anhand zweier Schwellenwerte können dann die eruierten Kollokationen in nichtsignifikante, mäßig signifikante und hochsignifikante eingeteilt werden und einer unterschiedlichen weiteren Bearbeitung zugeführt werden. Die hochsignifikanten Kollokationen sind zu einem späteren Zeitpunkt der Analyse zu unterscheiden in Kollokationen, die lediglich auffällige lexikalische Festigkeit zeigen (darunter fallen u.a. Kollokationen im Sinne der Phraseologieforschung) und in solche, die zudem Idiomatizität zeigen.

Aus dem Bestand der mäßig signifikanten und hochsignifikanten Kollokationen ergeben sich die für die weitere Analyse relevanten Kollokanteme. Pro Kollokantem wird ein Kollokatspotential gebildet und hinsichtlich seiner durchschnittlichen Signifikanz charakteri-

siert. Diese Signifikanzwerte werden u.a. dazu benutzt, den Ähnlichkeitsgrad zwischen Kollokatspotentialen zu berechnen und die Kollokanteme mit hohem Ähnlichkeitgrad in je einer Kollokantemgruppe zusammenzufassen.

Der skizzierte Verfahrensweg macht auch deutlich, weshalb wir von *lexikalischen Netzen* sprechen: Jede aus den verwendeten Korpora eruierte Analyseeinheit fungiert als Kollokant und ist damit einem Kollokantem zugeordnet. Jedes Kollokantem wiederum ist mit einem Kollokatspotential verknüpft und zusätzlich Element einer Kollokantemgruppe. Jedes Element eines Kollokatspotentials ist ebenfalls mit einer sprachlichen Einheit zu identifizieren, die als Kollokantem fungiert, etc.

Im Ergebnis finden sich die Analyseeinheiten in zweifacher Weise charakterisiert. Zum einen ist ihnen ein Kollokatspotential zugeordnet, das all ihre relevanten syntagmatischen Nachbarn in konkreten Texten in einer Menge vereint. Damit beruhen Kollokantem-Kollokat-Beziehungen auf dem Prinzip der Kontiguität. Zum anderen sind die Analyseeinheiten selbst in Kollokantemgruppen zusammengefaßt. Die wechselseitigen Beziehungen der Elemente einer Kollokantemgruppe beruhen auf dem Prinzip der Ähnlichkeit. Entscheidend ist, daß sich die genannten Prinzipien Kontiguität und Ähnlichkeit nicht allein auf sprachinterne Gegebenheiten beschränken. Vielmehr erstrecken sie sich auch auf die Objekte und Objektkonstellationen, die mit den sprachlichen Einheiten, die in einer solchen Beziehung stehen, benannt werden. Somit stellen Kollokatspotentiale eine geeignete Basis zur Konzeption von Frames dar und Kollokantemgruppen zeigen paradigmatische Bedeutungsbeziehungen wie Hyperonym-Hyponym-Beziehungen und Kohyponymbeziehungen auf.

Aufgrund der Analogien, die zwischen sprachinternen und sprachexternen Gegebenheiten zu verzeichnen sind, kann ein lexikalisches Netz die in ihm enthaltenen Elemente (selbst Homographe und Phraseme) in eindeutiger Weise identifizieren. Genutzt werden kann ein solches Netz als Datenbasis zu vielen unterschiedlichen Zwecken. Es bietet sich zum einen als *Datenbasis für jedwede Art von Textgenerierungs- und Textanalyseaufgaben* an; zum anderen können die maschinellen Verfahren, die die Bewältigung dieser Aufgaben leisten sollen, ebenfalls auf dem KOLYNET-Analysemodell fußen. Außerdem eignen sich ausgewählte Teile des KOLYNET-Modells zur Entwicklung maschineller Verfahren für andere und/oder begrenztere Zielsetzungen. Unter anderem lassen sich so - um nur einige Beispiele aus dem Bereich der Lexikographie zu nennen - konsistente wörterbuchinterne Verweissysteme maschinell erstellen (vgl. 4), lexikographische Beschreibungsvokabulare identifizieren und elektronische Belegsammlungen organisieren.