

Rekapitulation für den 21. 07. 2008

Einleitung

Das Seminar begann mit einigen knappen Kommentaren darüber, was Sie bezüglich Aufbau und Inhalt erwartet. Wir wollen uns mit einigen der zentralen Methoden und Konstrukten beschäftigen, die für die moderne Linguistik bzw. Computerlinguistik relevant sind, wie z.B.

- Phrasenstrukturen,
- Merkmale,
- Unifikation,
- Parsing.

Diese Begriffe finden auf verschiedenen sprachlichen Ebenen Anwendung, also der Phonetik und Phonologie, der Morphologie, der Syntax usw. In diesen Bereichen führen sie eine von der Computerlinguistik unabhängige Existenz: sie spielen also generell in der Sprachwissenschaft eine entscheidende Rolle.

Wir werden diese Begriffe über syntaktische Fragestellungen einführen und uns dabei zunächst auf einfache, unmarkierte Aussagesätze konzentrieren. Bei deren Behandlung ist zu berücksichtigen, dass dieser Kurs komplett theorieneutral ist: Sie werden nicht in einer bestimmten Ausrichtung (HPSG, Kategorialgrammatik, System Functional Grammar etc.) ausgebildet sondern sollen das Rüstzeug erhalten, das es Ihnen dann ermöglicht, sich diese Theorien später, in eigenen Kursen oder im Selbststudium, problemlos anzueignen.

Es geht uns also im Wesentlichen darum, die für

- diverse konkrete Anwendungen im Bereich Computerlinguistik
- das Verständnis moderner Ansätze im Bereich Grammatik / Sprachverarbeitung

relevanten linguistischen Grundkenntnisse zu vermitteln. Zu diesen Grundkenntnissen gehören bestimmte Methoden und Verfahren, eine große Menge an theoretischen Konstrukten und Termini, und – ganz wichtig – ein Verständnis dafür, was eigentlich die zentralen Fragestellungen sind, mit denen man sich auseinandersetzt.

Der Einstieg

Die erste Übung griff genau diesen letzten Punkt auf:

Syntax: ein informeller Einstieg

Gegeben sei die folgende Wortkette:
milben, kopfkissen, im, leben

❖ Wieviele Umstellungen (sog. *Permutationen*) sind für diese Wortkette möglich?
24 ($1 \times 2 \times 3 \times 4$)

❖ Wieviele von diesen Permutationen können Ihrer Ansicht nach als grammatische Sätze des Deutschen gelten?
6

❖ Wie könnte ein sehr einfaches System aussehen, das nur die grammatischen, nicht die ungrammatischen Sätze ausgibt?
Jegliche Permutationen dieser Kette liefert genau dann einen grammatischen Satz des Deutschen, wenn die Elemente 'im' und 'kopfkissen' in dieser Reihenfolge eine Einheit bilden, in die nichts eingefügt wird.

S. Hackmack

Uni Bremen / FB 10

Diese Übung hatte verschiedene Funktionen. Sie sollte verdeutlichen, dass

1. Sie – als kompetente Sprecher des Deutschen – durchaus über das linguistische Wissen verfügen, das man braucht, um die Daten zu beschreiben, dieses jedoch nicht unbedingt parat haben.
Genau um dieses Wissen geht es in der Linguistik bzw. der Computerlinguistik, wo es transparent gemacht und modelliert werden soll
2. man bei der Untersuchung sprachlicher Daten systematisch vorgehen muss und sich, bevor man anfängt, mit irgendwelchen Klassen oder Merkmalen wie NP oder Subjekt zu hantieren, zunächst einmal die zu beschreibenden Daten genau ansieht und analysiert.
Der Einsatz der entsprechenden Fachterminologie kommt erst dann zum Zuge, wenn es darum geht, diese Analysen zu generalisieren

Ferner demonstriert dieses Beispiel

1. eine der grundlegenden Kernfragen der Syntax, nämlich die (eigentlich ganz harmlose) Frage: Warum gilt die Wortkette XYZ als grammatisch, die Wortkette YZX dagegen nicht, und wie kann man das beschreiben?
2. zwei Problemstellungen, mit denen sich die Syntax beschäftigt, nämlich
 - a. die Beziehung 'bedeutungsgleicher', aber unterschiedlich organisierter Satztypen zueinander, vgl. *Milben leben im Kopfkissen* vs *Im Kopfkissen leben Milben*
 - b. die interne Struktur eines Satzes (*'im Kopfkissen* ist eine Einheit')

Der letzte Punkt dieser Aufzählung lieferte uns das Stichwort für den nächsten Abschnitt, in dem wir uns mit dem Strukturbegriff auseinandersetzen.

Struktur

An einem informellen Beispiel (dem Stammbaum der Simpsons) wurden dabei zunächst die Begriffe 'Struktur' und 'hierarchische Struktur' erläutert; anschließend wurden die diversen Repräsentationsformalismen demonstriert, die eingesetzt werden können, um Strukturen darzustellen:

- Ganz allgemein bezieht sich der Begriff 'Struktur' auf eine Menge von Elementen und den zwischen diesen Elementen bestehenden Relationen.
- Eine Klassifikation der Elemente (im Simpson-Beispiel z.B. die Klassen mann(X) und frau(x)) sowie der Relationen (z.B. vater(X,Y) und tante(X,Y)) erlaubt eine Generalisierung der beobachteten Relationen bzw. der Struktur.
- Eine Menge von asymmetrischen Relationen resultiert in einer hierarchischen Struktur.
- Für hierarchische Strukturen gibt es diverse Repräsentationsformen, ua. Klammerausdrücke, Tabellen oder Baumgraphen.

Als nächstes stand einer der Kernbegriffe der modernen Sprachwissenschaft auf dem Programm; nämlich Konstituenz. Erneut wurde dieser Begriff informell am Beispiel eines nicht-sprachlichen Gegenstandes (genau – wieder die Simpsons) eingeführt.

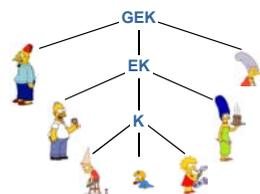
Konstituenz

- Ganz allgemein bezieht sich der Begriff 'Konstituente' auf eine Einheit von Elementen, die (ggf.) Teil einer größeren Einheit ist. Diese Einheit kann aus nur einem Element bestehen.
- Für die Darstellung von Konstituentenstrukturen gibt es verschiedene Repräsentationsformen, z.B. Klammersausdrücke, Tabellen oder Baumgraphen
- Man unterscheidet zwischen *unmittelbaren* und *nicht-unmittelbaren* Konstituenten. Eine Konstituente, die keine weitere Konstituenten dominiert (die also keine *Töchter* hat), wird *terminale Konstituente* genannt.

Konstituenz

Konstituentenstruktur als Baumgraph II

Diese Struktur ist wie folgt zu 'lesen':



Die Konstituente GEK setzt sich zusammen aus den Konstituenten Grampa Homer, EK und Grandma Bouvier.

Die Konstituente EK setzt sich zusammen aus den Konstituenten Homer, K & Marge. Die Konstituente K setzt sich zusammen aus den Konstituenten Bart, Maggie & Lisa.

Ein essentiell wichtiger Punkt war dabei der Umstand, dass, sobald einmal auf Grundlage bestimmter Daten (hier: Simpsons-Familienrelationen) eine Konstituentenstruktur erstellt wird, diese ggf. auch auf andere Daten (in unserem Beispiel die *Royal Family*) übertragen werden kann: die Struktur ist dann eine Art generelles Muster für diverse konkrete Realisierungen.

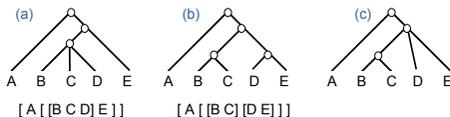
Auf diesen Punkt werden wir noch häufig zurückkommen

Diese Ausführungen lieferten die Grundlage für Übung 3:

Konstituenz

Beantworten Sie bitte die nachstehenden Fragen zu dem folgenden Klammersausdruck: [A [[B C] D E]]

- Das Syntagma BCD bildet eine Konstituente **falsch**
- Die Konstituente D ist eine terminale Konstituente **wahr**
- Das Syntagma ABCD bildet eine Konstituente **falsch**
- Die Konstituente BCDE hat vier Töchter **falsch**
- Die Konstituente BC ist unmittelbare Konstituente von ABCDE **falsch**
- Welcher der folgenden Ausdrücke ist/sind strukturell gleich mit dem oa. Ausdruck:
(a) [1[[23]45]] (b) [V[W[XY]Z] (c) [er [[den Hund] getreten hat]]
- Welchem der nachstehenden Bäume entspricht der Klammersausdruck? c
- Erstellen Sie für die beiden anderen Bäume Klammersausdrücke

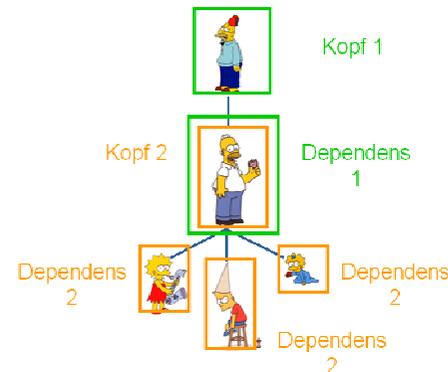


Als nächstes ging es um ein weiteres Kernkonstrukt der modernen Linguistik, nämlich den Begriff 'Dependenz'!

Dependenz

- Ganz allgemein bezieht sich der Begriff Dependenz auf eine zweistellige, asymmetrische Relation zwischen Elementen.
- In dieser Relation determiniert das eine Element, der sogenannte *Kopf*, das Auftreten, die Form oder allgemein das 'Verhalten' des anderen Elementes, des sogenannten *Dependens*. Der Kopf regiert sein Dependens.
- Für die Darstellung von Dependenzstrukturen gibt es verschiedene Repräsentationsformen, z.B. Baumgraphen oder linearisierte Formen. In allen Fällen gilt, dass Kopf und Dependens klar gekennzeichnet werden müssen - sei es über ihre vertikale Anordnung im Baumgraph oder entsprechend gekennzeichnete Kanten in der linearen Darstellung.

Als Beispiel für eine Dependenzstruktur leistete uns Familie Simpson erneut treue Dienste. Hier ging es wohlgernekt um die Form, nicht den Inhalt: wir haben ein intuitives Verständnis dafür, dass Grampa Homer; Homer wiederum jeweils Lisa, Bart und Maggie 'regiert' – ohne Grampa kein Homer, ohne Homer weder Lisa, Bart noch Maggie:



Diese Ausführungen lieferten die Grundlage für Übung 4, in der zudem die Tatsache deutlich wurde, dass in Dependenzstrukturen i.d.R. zwei Arten von Information kodiert sind, nämlich

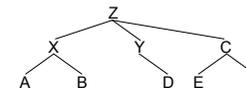
- Anordnungsrelationen (Linearität: X steht vor Y und hinter Z) und
- Hierarchie (X regiert Z, Y regiert X).

Diese Differenzierung (*linear precedence* vs *immediate dominance*) zieht sich wie ein roter Faden durch die Strukturbäume, die wir später erstellen werden.

Diese Ausführungen lieferten die Grundlage für Übung 4:

Dependenz

- Stellen Sie die folgende Information in Form eines Dependenzbaumes dar: Das Element A und B werden von X regiert. X selber ist Dependens von Z, das neben X noch Y und C regiert. Y regiert D, C regiert E und F. Die lineare Anordnung lautet A, X, B, Z, Y, D, E, C und F.



- Geben Sie dieselben Dependenzrelationen in linearisierter Form bei folgender Anordnung wieder: D, E, F, C, X, A, Y, B und Z



Im nächsten Teil ging es darum, diese beiden Konstrukte zusammenzubringen.

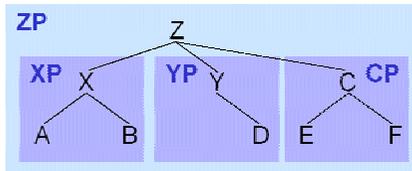
Dependenz, Konstituenz und Phrasen

Konstituentenstrukturen können aus Dependenzstrukturen nach folgendem Prinzip abgeleitet werden:

Ein Kopf plus die von ihm abhängigen Elemente, so vorhanden, bilden eine Konstituente.

Konstituenten werden i.d.R. nach ihrem Kopf benannt.

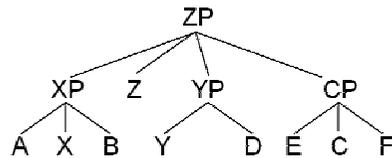
In der Syntax wird eine Konstituente 'Phrase' genannt. Ausgehend von der Annahme, dass in dem nachstehenden Dependenzbaum jeder nicht-terminale Knoten eine Phrase konstituiert, können wir nach dem o.a. Prinzip eine Z-Phrase (ZP), eine X-Phrase (XP), eine Y-Phrase (YP) und eine C-Phrase (CP) ableiten:



Annotierte Dependenzstruktur

In der annotierten Dependenzstruktur steckt also auch Information über die Konstituenz des analysierten Syntagmas – kurz gesagt: über dessen phrasale Struktur.

Diese wird allerdings gemeinhin etwas anders notiert, nämlich wie folgt:



Phrasenstruktur

Beide Strukturbäume stellen bezüglich Linearität und Hierarchie dieselbe Information dar. Sie unterscheiden sich darin, dass in der Phrasenstruktur das Konzept 'Kopf' nicht mehr explizit ausgedrückt ist, sondern nur noch implizit über die Benennung der Phrasen.

Diese Ausführungen lieferten die Grundlage für Übung 5:

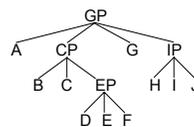
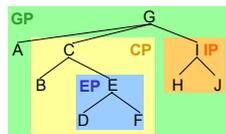
Dependenz, Konstituenz und Phrasen

1. Übersetzen Sie den folgenden Ausdruck in einen Dependenzbaum:



2. Markieren und benennen Sie in diesem Baum alle Phrasen. Es gelte: alle nicht-terminalen Knoten konstituieren eine Phrase.

3. Übersetzen Sie den annotierten Dependenzbaum in einen PS-Baum



Dependenzstrukturen können also eingesetzt werden, um Phrasen zu definieren: So gesehen kann zwischen Dependenz und Konstituenz ein enger Zusammenhang ausgemacht werden: Dependenzstrukturen lassen sich in Konstituentenstrukturen überführen.

Wissenschaftsgeschichtlich und auch formal betrachtet sind Dependenz und Konstituenz aber zwei voneinander unabhängige Konzepte.

Eine Dependenzstruktur ist eine Menge hierarchisch geordneter binärer Relationen zwischen (jeweils zwei) Wortformen. Eine Dependenzgrammatik verallgemeinert diese Relationen unter Zuhilfenahme der lexikalischen Kategorien. Das Konzept 'Kopf' ist von essentieller Bedeutung und in jeder Dependenzstruktur formal ausgewiesen.

Phrasenstrukturen dagegen sind Mengen hierarchisch geordneter Syntagmen. Eine 'traditionelle' Phrasenstrukturgrammatik verallgemeinert diese unter Berücksichtigung distributioneller Kriterien und Zuhilfenahme der lexikalischen Kategorien. Das Konzept 'Kopf' ist in ihnen nicht explizit ausgedrückt.

In der Entwicklung der Grammatik- oder Syntaxtheorie können sowohl Ansätze ausgemacht werden, die sich primär auf Dependenz als zentrales Konzept berufen, um syntaktische Regelmäßigkeiten auszudrücken, als auch solche, die den Schwerpunkt im Wesentlichen auf Konstituenz gelegt haben.

Syntaktische Regelmäßigkeiten werden in Dependenzgrammatiken über zweistellige Relationen zwischen Lexemklassen erfasst; in Phrasenstrukturgrammatiken über Klassen von Konstituenten.

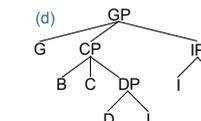
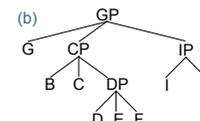
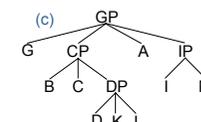
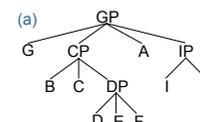
In moderneren Phrasenstrukturen bzw. Phrasenstruktur-grammatiken ist Information über Dependenz allerdings fest – dh. auch formal fest – integriert. Diese Wende vollzog sich ca. Anfang der siebziger Jahre des letzten Jh. mit Einführung der sogenannten X-bar Theorie. Wir werden auf diese noch zurückkommen, die genaue Ausdifferenzierung zwischen Dependenz- und PS-Grammatiken aber nicht weiter diskutieren.

Phrasenstrukturregeln

Im letzten Abschnitt des ersten Sitzungstages wurde schließlich demonstriert, wie aus derartigen PS-Bäumen Regeln abzuleiten sind. Da wir diese Tätigkeit noch x-mal im Kurs praktizieren werden, hier nur Übung 6 + dazugehörige Lösung:

Phrasenstrukturregeln

Leiten Sie PS-Regeln aus den folgenden PS-Bäumen ab:



Geschweifte Klammer stehen für Alternativen; runde Klammern schließen fakultative Elemente ein. In beiden Fällen handelt es sich um vereinfachende Schreibweisen.

GP → G CP (A) IP

CP → B C DP

IP → I { J }
 { M }

DP → D { E F }
 { (K) L }

Wie Ihnen vielleicht aufgefallen ist, hat kaum etwas von dem, was wir diskutierten, direkt mit Sprache bzw. Syntax zu tun.

Genau darum geht es am nächsten Sitzungstag: wir wollen dann, aufbauend auf den Erkenntnissen über formale Ausgestaltung von Strukturen, dieses Wissen mit sprachlichen Daten in Zusammenhang bringen.