Begriffsbestimmung: Was sind syntaktische Strukturen?

In seinem grundlegenden Werk „Syntactic Structures" aus dem Jahr 1957 definiert Noam Chomsky, nach dem auch die Unterscheidung verschiedener Sprachkomplexitäten benannt ist *(Chomsky-Hierarchie)*, Syntax als „the study of principles and processes by which sentences are constructed in particular languages" (Syntactic Structures, [2] S.1). Gegenstand syntaktischer Repräsentationen sind also die Wörter bzw. Wortformen einer Sprache und deren Kombination zu Sätzen.

Beispiel: Grammatische und ungrammatische Abfolge von Wörtern

Gegeben sei eine Menge von Wortformen, z.B. die Menge {Sonne, die, warm, Leipzig, in, scheint}. Welche Kombinationen daraus sind grammatikalische oder sinnvolle Sätze? Offenbar zählt (zumindest fürs Deutsche) die Reihenfolge, denn Folgen wie „die Leipzig scheint warm in Sonne" oder „warm die in Sonne scheint Leipzig" sind keine akzeptablen Sätze des Deutschen, wohl aber Folgen wie „die Sonne scheint in Leipzig warm" oder „warm scheint die Sonne in Leipzig".

Anders als bei einer formalen Sprache, z.B. einer Programmiersprache, ist allerdings für eine natürliche Sprache nicht immer eindeutig definiert, welche Kombinationen von Wortformen zulässig sind. Außerdem können Kombinationen von Wortformen strukturell mehrdeutig sein wie z.B. der Satz „Johann sieht den Mann mit dem Fernrohr". Aufgabe einer syntaktischen Repräsentation ist es also, nicht nur Regeln oder Modelle anzugeben, nach denen sich die Menge der grammatikalisch gültigen Sätze bestimmen lassen, sondern insbesondere auch die Prinzipien zu verdeutlichen, nach denen Kombinationen von Wortformen eine syntaktische Repräsentation zugewiesen wird, insbesondere auch strukturell mehrdeutigen Sätzen.

Im Folgenden stellen wir drei Ansätze syntaktischer Repräsentationen vor: Die Konstituenten-Struktur und die Dependenz-Struktur als regelbasierte Ansätze sowie das sog. probabilistische Parsen als sprachstatistischer Ansatz. Es handelt sich dabei um eine repräsentative Auswahl von Ansätzen, die beim Text Mining besonders häufig verwendet werden. Eine gute Übersicht syntaktischer Theorien aus Sicht der Linguistik und in historischer Perspektive findet sich in (Rauh 2010, [5]). Für die linguistische Theoriebildung sind dabei Fragen nach der Rolle des Lexikons bei der syntaktischen Repräsentation sowie die Abgrenzung der syntaktischen Ebene gegenüber der morphologischen und semantischen Ebene von zentraler Bedeutung. Für unsere Zwecke können wir diese Aspekte jedoch vernachlässigen und uns nur auf die zentralen Konzepte konzentrieren.

Konstituenten-Syntax

Im Sinne der sog. amerikanischen Strukturalisten Bloomfield ([1]) und Harris ([7]) sind Konstituenten auf der syntaktischen Ebene die unmittelbaren Bausteine von Sätzen, also Wortformen und Kombinationen von Wortformen, sog. Phrasen. Empirisch lassen sich Konstituenten durch eine Reihe von Tests bestimmen, welche hinreichende Bedingungen für das Vorliegen einer Konstitutente definieren, wie z.B. der Verschiebetest und der Ersetzungstest (vgl. Grewendorf et.al. [6]):

Verschiebetest

Gegeben sei der syntaktisch gültige Satz „die Sonne scheint in Leipzig warm". Alle Wortformen bzw. Gruppen von Wortformen, die sich in diesem Satz verschieben lassen, ohne dass ein ungrammatischer Satz entsteht, sind Konstituenten.

Nicht verschieben lassen sich der Artikel „die" oder das Nomen „Sonne", wohl aber die Kombination beider Wortformen als sog. Nominalphrase, wie die folgenden syntaktisch gültigen Verschiebungen zeigen:

„scheint die Sonne in Leipzig warm" ,

„scheint in Leipzig die Sonne warm",

„scheint in Leipzig warm die Sonne"

Nach dem gleichen Prinzip lassen sich für diesen Beispielsatz die Phrasen „scheint“, „in Leipzig“ und „warm“ bestimmen.

Beim Ersetzungstest wird geprüft, ob eine Wortform oder Phrase durch eine andere ersetzt werden kann, ohne dass ein ungrammatischer Satz entsteht. Ist dies der Fall, ist die ersetzte Wortform oder Phrase eine Konstituente.

Ersetzungstest

Für den gegebenen Beispielsatz bestätigen z.B. die Ersetzung der Phrase „in Leipzig“ durch die Phrase „in Berlin“ oder der Wortform „warm“ durch „heiß“ das Ergebnis des Verschiebetests.

In der sog. Phrasen-Struktur-Grammatik (PSG), wie sie Chomsky erstmals in dem bereits erwähnten Buch Syntactic Structures vorgeschlagen hat, ist eine Konstituente ein Wort oder eine Phrase, die von einer syntaktischen Regel als Einheit behandelt wird. Der Rahmen für die syntaktischen Regeln ist dabei eine kontextfreie Ersetzungsgrammatik (also in der Chomsky-Hierarchie eine Typ-2 Grammatik) mit allen Wortformen einer Sprache als terminalen Symbolen, den syntaktischen Kategorien und Phrasen als nichtterminalen Symbolen und den syntaktischen Regeln als Produktionsregeln. Grundlegende Konstituenten sind dabei die syntaktischen Kategorien *Nomen (N), Verb (V), Adjektiv/Adverb (A), Präposition (P)* und *Artikel (Art),* welche als Nichtterminale über das Lexikon direkt den terminalen Wortformen zugewiesen werden, sowie die nichtterminalen Phrasen *NP, VP, AP* und *PP*. Syntaktische Strukturen werden in Form eines Syntaxbaums beschrieben, mit dem auch syntaktische Mehrdeutigkeiten dargestellt werden können.

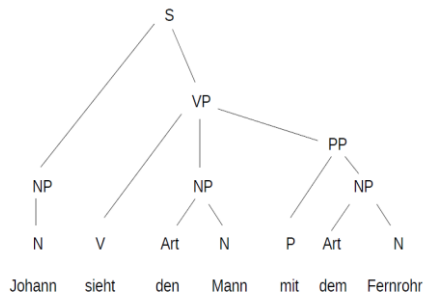
Syntaktische Mehrdeutigkeiten

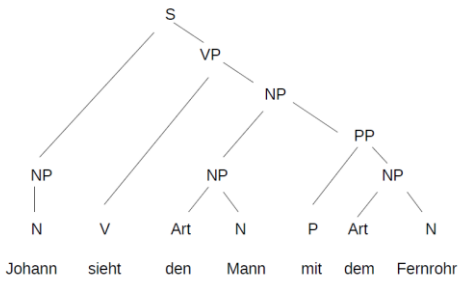
Der Satz „Johann sieht den Mann mit dem Fernrohr“ ist strukturell mehrdeutig, je nachdem, wer das Fernrohr hat (Johann oder der Mann). Nehmen wir an, den Wortformen in diesem Satz seien die folgenden syntaktischen Kategorien (KAT) zugeordnet (N=Nomen, V=Verb, Art=Artikel, P=Präposition):

KAT(Johann) = N, KAT(sieht) = V, KAT(den) = Art, KAT(Mann) = N,

KAT(mit) = P, KAT(dem) = Art, KAT(Fernrohr) = N

Die beiden Lesarten lassen sich dann durch die folgenden beiden Bäume (mit dem Startsymbol S) darstellen:





Schon einfache Verschiebungen von Phrasen lassen sich in der Phrasen-Struktur-Grammatik schwer darstellen. Chomsky hat deshalb in dem bereits erwähnten Buch Syntactic Structures ([2]) sog. Transformationen als Ergänzung der kontextfreien Regeln vorgeschlagen, mit denen u.a. Probleme der freien Wortstellung und Morphologie behandelt werden sollen. Transformationen dienen im Rahmen der Transformationsgrammatik dazu, die beobachteten Wortfolgen (sog. Oberflächenstrukturen) durch Veränderungen und Verschiebungen von Phrasen aus sog. Tiefenstrukturen abzuleiten, welche mit Hilfe der kontextfreien PSG erzeugt worden sind (vgl. Aspects [3]).

Transformation

Der Satz der Oberflächenstruktur „Mit dem Fernrohr sieht Johann den Mann" lässt sich aus der Tiefenstruktur „Johann sieht den Mann mit dem Fernrohr" durch folgende Transformation erzeugen:

> **Strukturbeschreibung**: 1 (Johann [N]) 2 (sieht [V]) 3 (den Mann [NP]) 4 (mit dem Fernrohr [PP])
>
> **Strukturveränderung**: 1 [N] 2 [V] 3 [NP] 4 [PP] ==> 4 [PP] 2 [V] 1 [N] 3 [NP]

Ein besonderes Problem für die PSG ist die Darstellung sog. thematischer Rollen, d.h. vom Verb ausgehende Anforderungen an die Anzahl und Art der ergänzenden Phrasen im Satz (vgl. Rauh [13]). Grundlage dieses Konzepts ist die Beobachtung, dass Wortformen nicht beliebig miteinander kombiniert werden können, sondern dabei auch Restriktionen zu beachten sind, die syntaktische und semantische Aspekte miteinander verbinden.

Syntaktische Restriktionen

Im Satz „Johann sieht den Mann mit dem Fernrohr" kann die Wortform „sieht" durch die Wortform „sucht" ersetzt werden, wobei die strukturelle Mehrdeutigkeit erhalten bleibt. Auch kann für beide Wortformen die Ergänzung „mit dem Fernrohr" weggelassen werden („Johann sieht den Mann"/"Johann sucht den Mann"), anders als die Wortform „sieht" kann aber die Wortform „sucht" nicht nur alleine mit dem Nomen stehen („Johann sucht" geht nicht, „sucht" hat keine intransitive Lesart).

Beide Verbformen können darüber hinaus nur durch Nomina ergänzt werden, die semantisch passen, also „Informatik sieht den Mann mit dem Fernrohr" geht genau so wenig wie „Johann sieht die Informatik mit dem Fernrohr".

Als syntaktische Strukturen, die mit einer kontextfreien Chomsky-Grammatik dargestellt werden können, eignen sich Konstituentenstrukturen besonders gut für die automatische Syntaxanalyse natürlichsprachlicher Sätze (in der Fachliteratur auch als *Parsing* bezeichnet). Hierzu gibt es umfangreiche Darstellungen, auf die in diesem Rahmen nicht weiter eingegangen werden soll (vgl. Jurafsky und Martin [10]). Standardprogramme für die automatische Syntaxanalyse von Sätzen in Konstituenten finden sich z.B. im Rahmen der Forschungsinfrastruktur CLARIN (https://weblicht.sfs.uni-tuebingen.de/weblicht/demo/)

Konstituenten bilden auch meist die Grundlage für **POS-Tags**. Darunter versteht man die Zuordnung von Wortformen und Satzzeichen eines Textes zu Wortarten (englisch: *part of speech*). Eine häufig verwendete Liste von POS-Tags fürs Deutsche ist das sog. Stuttgart-Tübingen Tag Set, STTS [14].

Für das Text Mining sind Konstituentenstrukturen relevant, weil sie mit dem Phrasen-Konzept eine wichtige Möglichkeit aufzeigen, wie Wortformen verschiedener Grundkategorien wie Nomen, Verb, Adjektiv und Präposition miteinander kombiniert werden können. Die Bestimmung von Phrasen wird insbesondere für die automatische Extraktion von Relationen benötigt. Die Idee der Transformationen bildet darüber hinaus eine wichtige Grundlage von Dialogsystemen (vgl. Weizenbaum [17]).

Dependenzen

Ein alternativer Ansatz zu Konstituentenstrukturen im Rahmen des regelbasierten Paradigmas sind die auf Lucien Tesnière zurückgehenden Dependenzstrukturen (*Éléments de syntaxe structurale* [15]). Während bei der syntaktischen Repräsentation von Konstituentenstrukturen die Darstellung der *Abfolge* von Wortformen im Mittelpunkt steht, stellen Dependenzstrukturen die zwischen den Wortformen eines Satzes bestehenden Abhängigkeiten *(Dependenzen)* dar. Dabei wird angenommen, dass jede Wortform in einem Satz eine, und nur eine, übergeordnete Wortform hat, von der sie abhängt. Ein wesentlicher Test für das Vorliegen einer Dependenz ist deshalb der Löschtest: Verliert eine Wortform nach dem Löschen einer anderen Wortform in einem Satz ihren Sinn, dann ist die gelöschte Wortform im Dependenzbaum der anderen Wortform übergeordnet.

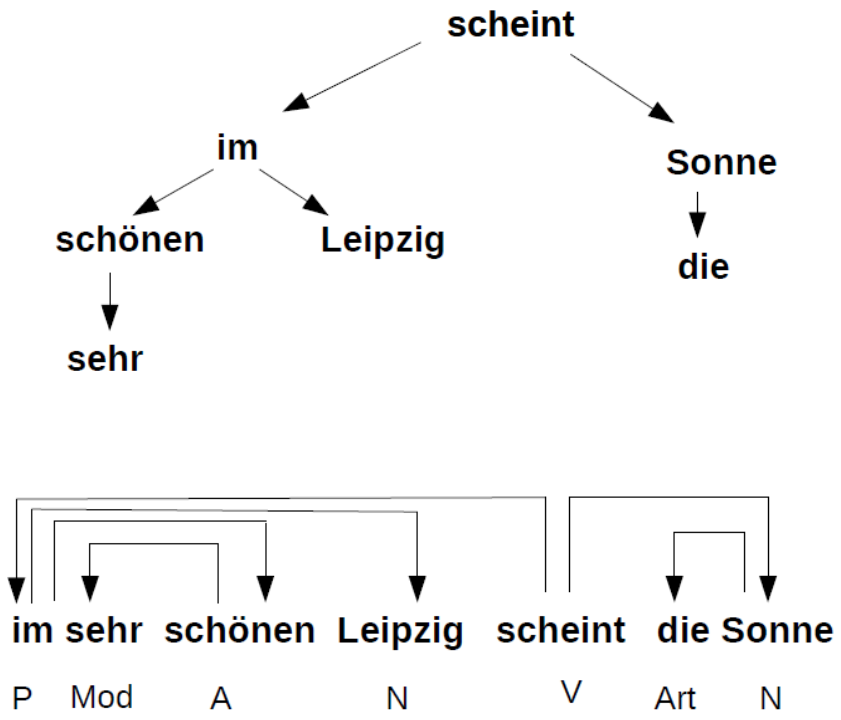
Dependenzstruktur

Betrachten wir den Satz „Die Sonne scheint im sehr schönen Leipzig". Wenn wir „schönen" löschen, müssen wir auch „sehr" löschen, weil „sehr" in der Folge von Wortformen „Die Sonne scheint im sehr Leipzig" keinen Sinn ergeben. Die Wortform „sehr" ist also abhängig von der übergeordneten Wortform „schönen".

Die Abhängigkeiten zwischen den Wortformen lassen sich ähnlich der Konstituentenstruktur in Form eines Baumes darstellen, wobei jedoch anders als in der Konstituentenstruktur-Syntax jeder Knoten im Baum einer terminalen Wortform entspricht. Als Wurzel einer Dependenzbaum-Struktur wird dabei immer das Verb (und nicht eine abstrakte Kategorie *Satz*) angesetzt. Ähnlich der Konstituentenstruktur-Syntax können dabei auch einzelne Knoten benannt werden, etwa mit den Grundkategorien N, V, A, P, Art und Mod.

Dependenzgraph

Der nachfolgende Baum stellt die Dependenz-Struktur des Beispielsatzes „Im sehr schönen Leipzig scheint die Sonne" dar:



Dependenzstrukturen erlauben eine sehr flexible und effektive Repräsentation syntaktischer Zusammenhänge. Für das Text Mining sind sie von besonderem Interesse, weil sie neben einer Benennung von Knoten auch eine Benennung von Kanten im Strukturbaum durch mehr inhaltlich-funktionalen Kategorien erlauben, wie beispielsweise Subjekt und Objekt eines Satzes. Standardprogramme für das Parsen von Sätzen in Dependenzen finden sich wieder z.B. im Rahmen der Forschungsinfrastruktur CLARIN (https://weblicht.sfs.uni-tuebingen.de/weblicht/demo/) sowie in dem von Google vorangetriebenen Projekt *SyntaxNet* (https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html).

Probabilistisches Parsen

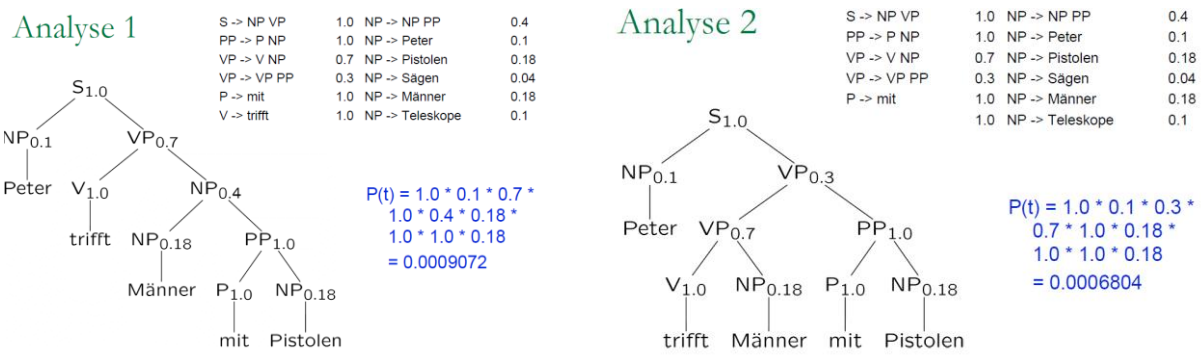
Im *regelbasierten* Paradigma wird für die Tokens einer natürlichen Sprache L eine *Grammatik*

angegeben, welche die – möglicherweise mehrdeutige – syntaktische Struktur einer Folge von Tokens in L beschreibt. Damit lässt sich auch bestimmen, ob eine Folge von Tokens in der Sprache L zulässig ist, also z.B. eine Phrase oder ein Satz ist (wobei mit der syntaktischen Korrektheit keineswegs sichergestellt ist, dass der zulässige Ausdruck auch semantisch sinnvoll ist). Demgegenüber wird beim *statistischen* Ansatz die *Wahrscheinlichkeit* bestimmt, dass eine Folge von Tokens in der Sprache L zulässig ist. Grundlage der Berechnung ist das Auftreten von Wortformkombinationen in einem Korpus. Damit werden sowohl syntaktische als auch semantische Aspekte berücksichtigt.

Eine *probabilistische* kontextfreie Grammatik (PCFG) ist eine Grammatik, in der jede Regel mit einer Wahrscheinlichkeit versehen ist, wobei die Summe der Wahrscheinlichkeiten aller Regeln mit demselben Symbol auf der linken Seite 1 betragen muss. Die Wahrscheinlichkeit einer Zerlegung ist dann das Produkt der Wahrscheinlichkeiten der Regeln, die während des Parsens angewandt werden (vgl. Manning und Schütze [11]):

Beispielableitung „Peter trifft Männer mit Pistolen"

Als Grundlage der syntaktischen Analyse verwenden wir wieder eine Phrasenstrukturgrammatik, wobei jedoch jede Regel mit einer Wahrscheinlichkeit für ihre Anwendung versehen ist.



Wie bei der Phrasenstrukturgrammatik erhalten wir mit den zwei abgeleiteten Bäumen eine Repräsentation der syntaktischen Mehrdeutigkeit dieses Satzes. Im Unterschied zur Phrasenstrukturgrammatik unterscheiden sich die beiden Ableitungen aber deutlich in den ihnen zugewiesenen Wahrscheinlichkeiten.

Für die Berechnung der Regelwahrscheinlichkeiten in einer PCFG kann ein syntaktisch analysierter Korpus wie z.B. der TIGER-Korpus fürs Deutsche [16] verwendet werden oder eine Baumdatenbank wie NEGRA ([12]), in der für alle Regeln bereits die Regelwahrscheinlichkeiten mit angegeben sind.

Literatur

[1] Bloomfield, Leonard, Language, University of Chicago Press 1984

[2] Chomsky, Noam, Syntactic Structures, Mouton 1957, Nachdruck bei Mouton - de Gruyter 2009

[3] Chomsky, Noam, Aspects of the Theory of Syntax, Cambridge, Massachusetts: MIT Press 1965

[4] Andrew Carnie (2010): *Constituent Structure*, 2nd edition, Oxford University Press

[5] Good, I. J. (1953): The Population Frequencies of Species and the Estimation of Population Prameters. Biometrika, Vol. 40, S. 237-264, Dezember 1953.

[6] Grewendorf, Hamm, Sternefeld, Linguistisches Wissen, Suhrkamp 2008

[7] Harris, Z. S.: Mathematical Structures of Language. Wiley, New York. 1968

[8] Jelinek, Fred; Mercer, Robert L. (1985): Probability distribution estimation from sparse data. IBM Technical Disclosure Bulletin, Vol. 28, S. 2591-2594.

[9] F. Jelinek, J. D. Lafferty, and R. L. Mercer. Basic methods of probabilistic context free grammars. In Pietro LafaceMori and

Renato Di, editors, Speech Recognition and Understanding. Recent Advances, Trends, and Applications, volume F75 of NATO Advanced Study Institute, pages 345–360. Springer Verlag, Berlin, 1992.

[10] Daniel Jurafsky, James H. Martin, Speech and Language Processing, Pearson/Prentice Hall 2009

[11] Manning, Christopher D.; Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. Cambridge/Massachusetts: MIT Press.

[12] NEGRA Baumdatenbank, http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html

[13] Gisa Rauh, Syntactic Categories, Oxford 2010

[14] STTS, http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

[15] Lucien Tesnière: *Éléments de syntaxe structurale* 1959

[16] TIGER Korpus https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger/

[17] Weizenbaum, Joseph, ELIZA A Computer Program For the Study of Natural Language Communication Between Man And Machine, Comnunications of the ACM, Volume 9 / Number 1, 1966.