Strukturalismus

Was ist die Grundidee des Strukturalismus?

Die Berücksichtigung der für eine Sprache typischen Gesetzmäßigkeiten auf allen ihren linguistischen Ebenen ist für die automatische Ermittlung von inhaltlichen Zusammenhängen aus Texten eine wesentliche Voraussetzung. Als theoretische Grundlage für die Entdeckung und Beschreibung linguistischer Strukturen beziehen wir uns im Folgenden auf den linguistischen *Strukturalismus*. Damit wird eine Richtung der Linguistik bezeichnet, die auf den Schweizer Linguisten Ferdinand de Saussure und seine posthum erschienenen Vorlesungsreihe „Einführung in die allgemeine Sprachwissenschaft" (Saussure 1916/1966, [1]) zurückgeht. Grundgedanke dieses Ansatzes ist die Beobachtung, dass Zeichenketten, wie wir sie bereits bei der Beschreibung der linguistischen Ebenen in Texten kennengelernt haben, – also *Buchstaben, Morpheme, Wortformen, Phrasen* und *Sätze* – immer nur in Bezug auf jeweils andere Zeichenketten dieser Ebene einen Sinn bzw. eine Funktion haben (vgl. Saussure 1916/2001, [1], S. 145).

> **Beispiel**
>
> Ein orthographisches Muster, wie z.B. „A" oder „H", hat eine Funktion als Buchstabe nur in einem Alphabet. Wählen wir als Bezugsrahmen das lateinische Alphabet für das Deutsche und das kyrillische Alphabet fürs Russische, dann hat zwar das Muster „A" jeweils die gleiche Aussprache, nicht aber das Muster „H", das im Russischen als „N" (wie in „Namen") gesprochen wird.

Die Beschreibung von linguistischen Strukturen in Texten benutzt dabei im Sinne des Strukturalismus zwei grundlegenden Dimensionen:

1. **Gemeinsames Auftreten** - Welche Zeichen bzw. Zeichenketten treten gemeinsam auf?
2. **Ähnliche Kontexte** - Welche Zeichen bzw. Zeichenketten treten in ähnlichen Kontexten auf?

In der Tradition des Strukturalismus wird die erste Dimension auch als **syntagmatische**, die zweite als **paradigmatische** bezeichnet.

Aus der Betrachtung beider Dimensionen können wir für das Text Mining zwei Aufgaben ableiten:

1. ausgehend von einem **Wort**: Vorhersage eines Kontextes und
2. ausgehend von einem **Kontext**: Vorhersage eines Wortes.

Im Folgenden sollen diese Dimensionen aufbauend auf dem Begriff des *Kontexts* genauer ausgeführt werden.

Kontexte

Für alle Texte einer Sprache L bezeichne die *linguistische Ebene* $e \in E$ (mit $E = \{Buchstaben, Morpheme, Wortformen, Phrasen, Sätze\}$) eine abzählbare Menge von gültigen Zeichen bzw. Zeichenketten A aus e. Die Elemente aus A bezeichnen wir als *Beispiele* für e aus L.

> **Beispiel**
>
> Das Alphabet fürs Deutsche umfasst als Beispiele die einzelnen Buchstaben, in geordneter Folge also „a", „b", „c" usw.
>
> Die Ebene der Morpheme umfasst bedeutungstragende Morpheme wie „geh" oder „lauf", aber auch Funktionsmorpheme wie „st" (in „geh-st"). Die Beispiele für Morpheme im Deutschen können in Form einer Liste angegeben werden.
>
> Für die Ebene der Wortformen und Sätze finden sich sortierte Listen von Beispielen in den Referenzkorpora deu_news_2015_10K-words.txt sowie deu_news_2015_10K-sentences.txt

Wesentlich ist die Beobachtung, dass die *Instanziierung* einer linguistischen Ebene dabei immer nur die zulässigen Beispiele der gewählten Ebene umfassen, also beispielsweise Listen von Wortformen. Dabei werden an diesen Listen aber auch typische Gesetzmäßigkeiten der darunter liegenden Ebene deutlich. So lassen sich z.B. aus einer Liste von Wortformen des Deutschen durch regelbasierte oder statistische Verfahren typische Buchstabenkombinationen oder morphologische Muster fürs Deutsche erkennen. Die Idee, für die Entdeckung und Beschreibung linguistischer Strukturen den *Kontext* der jeweiligen Zeichenketten auszuwerten, können wir also in einem ersten Schritt insofern präzisieren, als dass stets für den *Kontext* der zulässigen Beispiele die jeweils darüber liegende Ebene bzw. Ebenen zu berücksichtigen ist bzw. sind.

Die Beschreibung linguistischer Strukturen im Sinne des Strukturalismus kann nunmehr anhand des Begriffs des *lokalen* und *globalen Kontexts* von Zeichenketten erfolgen. Wir definieren zunächst den Begriff des *lokalen Kontexts*.

Ein **lokaler Kontext** $K(a_i)$ eines Beispiels $a_i \in A$ aus einer linguistischen Ebene e enthält eine Menge von Zeichen bzw. Zeichenketten, mit denen a_i auf einer darüber liegenden Ebene $e+$ gemeinsam auftritt :

$$K(a_i) = \{a_1, a_2, ..., a_i, ... , a_n\} \setminus a_i$$

Beispiel lokaler Kontext

Gegeben sei die Ebene der Wortformen fürs Deutsche mit den folgenden Beispielen A1 bis A4 mit der darüber liegenden Ebene der zulässigen Sätze:

A_1 = {a_{11}= die, a_{12} = Sonne, a_{13} = scheint}

A_2 = {a_{21}= ein, a_{22} = Kind, a_{23} = lacht, a_{24} = fröhlich}

A_3 = {a_{31}= die, a_{32} = Kerze, a_{33} = scheint}

A_4 = {a_{41}= die, a_{42} = Sonne, a_{43} = lacht}

Der lokale Kontext für die Wortform „Sonne" im Beispiel (A1) ist dann

K_1 (a_{12} = Sonne) = { a_{11}= die, a_{13} = scheint}

Der **globale Kontext** $K_G(a_i)$ eines Beispiels $a_i \varepsilon A$ aus einer linguistischen Ebene e umfasst alle Zeichen bzw. Zeichenketten, mit denen a_i auf einer darüber liegenden Ebene $e+$ in L gemeinsam auftritt. Der globale Kontext eines Beispiels ist also die Vereinigungsmenge seiner lokalen Kontexte K_j:

$$K_G(a_i) = \cup_j (K_j(a_i))$$

Beispiel globaler Kontext

Für das obigen Beispiele A1 bis A4 mit der darüber liegenden Ebene der zulässigen Sätze ist der globale Kontext für die Wortform „scheint" die Vereinigung der Menge der lokalen Kontexte, welche die Wortform „scheint" enthalten:

KG (scheint) = K_1 (a_{13} = scheint) ∪ K3 (a_{33} = scheint)

= {a_{11}=a_{31} = die, a_{12} = Sonne, a_{32} = Kerze}

Für die Analyse des *gemeinsamen Auftretens* von Zeichen bzw. Zeichenketten auf einer Ebene sind einige Elemente des globalen Kontextes vermutlich relevanter als andere. Hierfür gibt es verschiedene Verfahren für die Berechnung der statistischen Signifikanz des *gemeinsamen Auftretens* von Zeichen bzw. Zeichenketten auf einer Ebene. Die Häufigkeit einzelner Wortformen dient dabei als Baseline, verschiedene Modellierungen der sprachstatistischen Gesetzmäßigkeiten *(Zipfverteilung)* führen im Ergebnis zu verschiedenen Maßen, welche einzelne Aspekte des *gemeinsamen Auftretens* von Zeichen bzw. Zeichenketten, z.B. wie *informativ* oder *überraschend* wir ein gemeinsames Auftreten bewerten, unterschiedlich gewichten. Wie das obige Beispiel verdeutlicht, lässt das *statistisch signifikante gemeinsame Auftreten* von Zeichen bzw. Zeichenketten auf einer Ebene dabei erwarten, dass zwischen den Zeichen bzw. Zeichenketten ein funktionaler und inhaltlicher Zusammenhang besteht. Dieser Zusammenhang gilt für alle linguistischen Ebenen.

Beispiele für statistisch signifikantes gemeinsames Auftreten von Zeichenketten

Auf der morphologischen Ebene finden wir etwa Gesetzmäßigkeiten der Art, dass bestimmte Pronomina stets bestimmte Endungen des darauffolgenden Verbs verlangen:

PERSON	PRONOMEN	VERB+ENDUNG
1	ich	geh+e, lach+e, …
2	du	geh+st, lach+st, …
3	er/sie/es	geh+t, lach+t, …

Auf der syntaktischen Ebene finden wir Muster, dass beispielsweise ein Nomen stets durch ein passendes Verb ergänzt werden kann und diese Kombination einen grammatikalisch vollständigen und korrekten Satz ergibt („Die Sonne scheint").

Auf der semantischen Ebene deutet das statistisch signifikante gemeinsame Auftreten von Zeichenketten auf inhaltliche Zusammenhänge hin, z. B. dass die Sonne scheint.

Zeichen bzw. Zeichenketten, die auf einer Ebene statistisch signifikant in *ähnlichen Kontexten* auftreten, haben meist eine grammatikalisch und inhaltlich ähnliche Funktion.

Beispiel

Für die obigen Beispiele A1 bis A4 mit der darüber liegenden Ebene der zulässigen Sätze ist der globale Kontext für die Wortform „Sonne" ähnlich dem globalen Kontext für die Wortform „Kerze":

KG (Sonne) = {a_{11}=a_{41} = die, a_{13} = scheint, a_{43} = lacht}

KG (Kerze) = {a_{31} = die, a_{33} = scheint}

Verschiedene Verfahren für die Berechnung der Ähnlichkeit von Kontexten behandeln wir in Kapitel 6. Neben einfachen Verfahren, welche die gemeinsame Anzahl von Zeichenketten auswerten, spielen dabei zunehmend auch vektorbasierte Verfahren eine Rolle. Dabei werden die statistisch signifikant mit einem Beispiel gemeinsam auftretenden Beispiele einer Ebene als Merkmale für die Elemente dieser Ebene aufgefasst und als Vektoren repräsentiert. Die Ähnlichkeit von Elementen einer Ebene kann dann als Abstand in dem entsprechenden Vektorraum berechnet werden.

Beispiel

Nehmen wir an, wir haben den Verwendungskontext der folgenden Wortformen aus einem großen Zeitungstextkorpus berechnet (also denjenigen Wortformen, die mit dem Ausgangswort statistisch signifikant gemeinsam Auftreten): Berlin, Paris, Deutschland, Frankreich. Werden die Verwendungskontexte dieser Wortformen als Vektoren repräsentiert, so befinden sich Berlin und Paris sowie Deutschland und Frankreich im Vektorraum nahe beieinander, was der semantischen Ähnlichkeit zwischen den beiden Hauptstädten und den Ländern entspricht.

Linguisten haben Ende des 19. Jahrhunderts damit begonnen, Verfahren für die Beschreibung von Sprachen zu entwickeln, deren Strukturen sich nicht aus den bekannten Strukturen der europäischen Sprachen ableiten lassen. Methodisch war dafür der **linguistische Strukturalismus** eine wesentliche Grundlage. Heute, mit der Verfügbarkeit sehr umfangreicher digitaler Textmengen für sehr viele Sprachen, Softwarearchitekturen und Algorithmen für sehr große Textmengen und Computersystemen, die effizient und kostengünstig deren Verarbeitung erlauben, können einige dieser Verfahren auch als Grundlage für die automatische Analyse von Texten genommen werden. Ein Programm, das automatisch aus einem Text wesentliche Inhaltsstrukturen ableiten soll, versteht von dem Text ja zunächst genau so wenig wie ein Linguist, der eine ihm unbekannte Sprache erlernen möchte.

Literatur

[1] Saussure, F. de, Grundfragen der allgemeinen Sprachwissenschaft, de Gruyter, Berlin ^{5}1966