Einführung in die Computerlinguistik
Grundkonzepte

Hinrich Schütze

Center for Information and Language Processing

2018-10-19

Die Grundfassung dieses Foliensatzes wurde von Dr. Benjamin Roth erstellt. Fehler und Mängel sind ausschließlich meine Verantwortung.

Outline

Outline

Natürliche Sprache

- In der Computerlinguistik beschreiben, modellieren, verarbeiten wir natürliche Sprache.
- Nicht: Programmiersprachen, Logiksprachen, Kunstsprachen (z.B. Boeing manuals)

Definition "Natürliche Sprache"?

- Gebärdensprache?
- Kommunikation unter Tieren (Menschenaffen, Delphine)?
- Latein, Sumerisch?
- Esperanto?
- Ein System von Zeichen (Wortschatz) und Regeln (Grammatik) zur Mitteilung von Bedeutungen?
- Hier kein Versuch der Definition ...
- Im Wesentlichen:
  Englisch, Deutsch und etwas 100 weitere Sprachen
- Typologisch sehr schlechte Abdeckung!

Outline

Begriffsklärung "Wort"

- Der Begriff "Wort" ist ungenau, wenn nicht weiter spezifiziert.
- Meinen wir das abstrakte Wort oder ein konkretes Vorkommen?
- Unterscheidungen:
  - Wortform vs. Lexem
  - Token vs. Type

Wortform vs. Lexem

- Wortform: flektierte Form eines Wortes, so wie sie im Text oder in geschriebener Sprache vorkommt.
  Beispiele: “sings”, “schönes”
- Ein Lexem ist eine Klasse lexikalisch äquivalenter Wortformen. Diese Wortformen repräsentieren das Lexem in verschiedenen Umgebungen.
  Beispiel: L1 = {“sing”, “sings”, “singing”, “sang”, “sung”}
- Oft wird auf ein Lexem mit seiner Zitierform Bezug genommen, z.B. Infinitiv oder erste Person Singular für Verben und Nominativ Singular für Nomen.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Gleichheitskriterien für Token

- Für die Anzahl der Types in einem Text macht es einen Unterschied, ob wir uns auf Wortformen oder Lexeme beziehen.
- Beispiel: “eine Rose ist eine Rose und viele Rosen ergeben einen Strauß”
- Wortformen:
  $\Rightarrow$ 11 Token, 9 Types
- Lexeme:
  $\Rightarrow$ 11 Token, 7 Types

Bestimmungskriterien für die Einheit "Wort"

- orthographisch / graphematisch
- phonologisch
- morphologisch
- morphosyntaktisch
- semantisch
- "Intuition"
- Literatur: Heringer, H.-J.: Morphologie. Paderborn 2009

Orthographisches Kriterium

- “Wörter sind sprachliche Einheiten, die als Folgen von Buchstaben zwischen Leerzeichen geschrieben werden.” aber:
- Sprachen ohne Buchstabenschrift
- weitere Trennzeichen
- abtrennbare Präfixe bei zusammengesetzten Verben
- zirkuläre Definition!

Phonologisches Kriterium

- "Wörter sind durch eine spezielle einheitliche Akzentstruktur gekennzeichnet, die sich von der entsprechender Wortgruppen/Phrasen unterscheidet."
- unterscheidbar: Wíenerwald vs. Wiener Wáld
  aber:
- präzisere Beschreibung der Intonationsmuster nötig

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

“klein”: Starke Adjektivflexion

	Singular			Plural
	Maskulinum	Neutrum	Femininum	
Nominativ	-er	-es	-e	-e
Akkusativ	-en			
Dativ	-em		-er	-en
Genitiv	-en			-er

klein+er	klein+e	klein+es	klein+e
klein+es/en	klein+er	klein+es/en	klein+er
klein+em	klein+er	klein+em	klein+en
klein+en	klein+e	klein+es	klein+e
klein+er+er	klein+er+e	klein+er+es	klein+er+e
klein+er+es/en	klein+er+er	klein+er+es/en	klein+er+er
klein+er+em	klein+er+er	klein+er+em	klein+er+en
klein+er+en	klein+er+e	klein+er+es	klein+er+e
klein+st+ er	klein+st+e	klein+st+ es	klein+st+e
klein+st+es/en	klein+st+er	klein+st+es/en	klein+st+er
klein+st+em	klein+st+er	klein+st+em	klein+st+en
klein+st+en	klein+st+e	klein+st+es	klein+st+e

Morphologische Kriterien

- a) “Ein morphologisches Wort ist eine grammatische Einheit, die nicht von Lexikoneinheiten unterbrochen werden kann.”
  aber:
- Im- und Export
- hin und her
- “Lexikoneinheit” $\rightarrow$ unbestimmt oder zirkuläre Definition
- b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.”
- aber:
- nicht flektierbare Wörter?!

Morphosyntaktisches Kriterium

- "Wörter sind die kleinsten sprachlichen Einheiten, die innerhalb des Satzes permutierbar sind."
  aber:
- syntaktische Regeln lassen oft keine Permutation zu
- "das kleine Haus" $\Rightarrow$ *"das Haus kleine"

Semantische Kriterien

- “[...] kleinste Einheiten des Inhalts oder der Bedeutung.”
- “[...] satzfähiges Lautsymbol mit der Eignung, ein Stück Wirklichkeit zu meinen.”
  aber:
- Funktionswörter, z.B. Partikel zu
- Idiome, mehrere Wörter für einen Begriff! z.B. roter Faden,
- Teilweise ist unklar, wie weit Bezeichner zerlegt werden sollten: Frankfurter-Straßennamen-Büchlein

Symptom der Schwierigkeit der Definition: Rechtschreibregeln

- Getrennt vs. zusammen schreiben
- Rad fahren vs. radfahren
- Das war nicht zu sehen vs. Das war nicht einzusehen

Kriterium: Intuition des Muttersprachlers

- Wort = durch Muttersprachler intuitiv erkennbare Basiseinheit des Lexikons
- Zirkulär!

Kriterium: Intuition des Muttersprachlers

- Dixon and Aikhenvald, 2007
  . . . the vast majority of languages spoken by small tribal groups . . . have a lexeme meaning ‘(proper) name’, but none have the meaning ‘word’.

Das Konzept “Wort”

- Der intuitive Begriff “Wort” ist kein eindeutig definiertes Konzept.
- Die Intuition wird mehr oder weniger gut anhand orthographischer/ graphemischer, phonologischer, morphologischer und semantischer Kriterien beschrieben.
- Viele Wörtern erfüllen alle Kriterien, es gibt aber immer Ausnahmen, die mit einigen Kriterien nicht übereinstimmen (Prototypen- oder Familienähnlichkeit).
- Wir wir sahen: teilweise ist unsere Intuition nicht eindeutig: “Rad fahren” vs “radfahren”
- Wortkonzept ist auch Kulturabhängig (bei gleicher Bedeutung und syntaktischer Funktion): “business trip” vs. “Dienstreise”
- Theorien, die das Konzept “Wort” unzweideutig definieren (wollen), weichen teils stark vom intuitiven Verständnis des Konzeptes ab.

Token vs. Type

- Token / Wortvorkommnis: Konkretes Vorkommen eines Wortes (z.B. vor oder nach einem anderen Token).
- Type / Worttyp:
  Ein Type bezeichnet eine Klasse von Token ...
  - ..., die nicht unterschieden werden
  - ..., die als Kopien wahrgenommen werden
  - ..., die gleich sind
- Gleichheit: verschiedene Kriterien der Unterscheidung, siehe nächste Folie
- "eine Rose ist eine Rose" $\Rightarrow$ 5 Token, 3 Types
- Verhältnis von Types zu Tokens (type-to-token ratio) ist eine wichtige Kennzahl zur Charakterisierung von Texten.

Übung

Wie viele Tokens und Types gibt es jeweils in folgenden Sätzen für (i) Wortformkriterium, (ii) Lexemkriterium?

1. Der Nachrichtensprecher versprach sich.
2. New York ist nicht die Hauptstadt der Vereinigten Staaten.
3. Er kauft gerne am Samstag ein.
4. Sie konnten weder vor- noch zurückgehen.
5. Hans war ganz aus dem Häuschen.

W:To	W:Ty	L:To	L:Ty	
5	5	5	5	Der Nachrichtensprecher versprach sich .
10	10	10	9	New York ist nicht die Hauptstadt der Vereinigten Staaten .
7	7	7	7	Er kauft gerne am Samstag ein .
7	7	7	7	Sie konnten weder vor- noch zurückgehen .
7	7	7	7	Hans war ganz aus dem Häuschen .

W:To	W:Ty	L:To	L:Ty	
5	5	5	5	Der Nachrichtensprecher versprach sich .
8	8	8	7	New_York ist nicht die Hauptstadt der Vereinigten_Staaten .
6	6	6	6	Er [ein]kauft gerne am Samstag _ .
7	7	7	7	Sie konnten weder vor- noch zurückgehen .
5	5	5	5	Hans war ganz aus_dem_Häuschen .

Outline

Syntagmatische und Paradigmatische Sprachachse

- syntagmatische Sprachachse:
  - Syntagma: Segmentierbare komplexe sprachliche Einheit; Ebene der Kombination.
  - Syntagmatische Relationen drücken die Beziehungen zwischen Aufeinanderfolgenden Teilen eines Satzes aus, z.B. von einem Zeichen (Token) zu einem anderen Zeichen in seinem Kontext.
    $\Rightarrow$ Grundlage zur Beschreibung der sprachlichen Struktur (Syntax)
- paradigmatische Sprachachse:
  - Paradigma: Menge von austauschbaren Zeichen bzw. Elementen derselben Kategorie; Ebene der Ersetzung.
  - Paradigmatische Relationen fassen sprachliche Einheiten aufgrund ihrer Ähnlichkeit in Kategorien (z.B. Wortarten) zusammen
  - Z.B. Beziehung von einem Zeichen (Lexem oder Wortform) zu anderen Zeichen des gleichen Paradigmas.
    $\Rightarrow$ Grundlage zur Beschreibung der sprachlichen Einheiten

Syntagmatische & Paradigmatische Sprachachse: Beispiel

Die Studentin	sitzt	in	der Vorlesung	↑
Ein Student	lernt	im	Seminar	paradigmatisch
Hans	liest	im	Hörsaal	↓

← syntagmatisch →

- Syntagmatische Relationen im Beispiel:
  - Hans ist Subjekt zu liest.
  - in der Vorlesung ist adverbiale Ergänzung zu sitzt
  - usw.
- Paradigmatische Relationen im Beispiel:
  - sitzt, lernt, liest sind Verben (3. Person Singular Präsens)
  - die Studentin, ein Student, Hans sind Nominalphrasen (Nominativ Singular)
  - usw.

Distribution eines Zeichens Z

- = Verteilung eines Zeichens Z
- Menge der Kontexte, in denen Z vorkommt
- z.B. “zwischen” kommt fast nur in Kontexten vor, deren rechter Teil eine Nominalphrase ist: “zwischen den Planzen”, “zwischen den Seiten”

Distributionsanalyse: Verfahren zur Ermittlung sprachlicher Strukturen (amerikanischer Strukturalismus)

1. Segmentierung in Einheiten (Intuition, morphologische Anhaltspunkte)
2. Überprüfen der Segmente und zusammenfassen in paradigmatische Klassen anhand der Ersetzungsprobe.
3. Finden von Syntagmatischen Relationen zwischen den paradigmatischen Klassen.

Wohlgeformtheit

- Ein sprachlicher Ausdruck A aus einer Sprache L heißt wohlgeformt, wenn er (laut Intuition der Sprecher von L) ein gültiger Ausdruck von L ist.
- Alternative: Ein sprachlicher Ausdruck A aus einer Sprache L heißt wohlgeformt, wenn er (laut Intuition der Sprecher von L) Sinn ergibt.
- Noam Chomsky (1957):
  Colorless green ideas sleep furiously.
  *Ideas green sleep colorless furiously
- nicht wohlgeformte Sätze (Ausdrücke) werden mit Stern gekennzeichnet

Deskriptivität vs. Präskriptivität

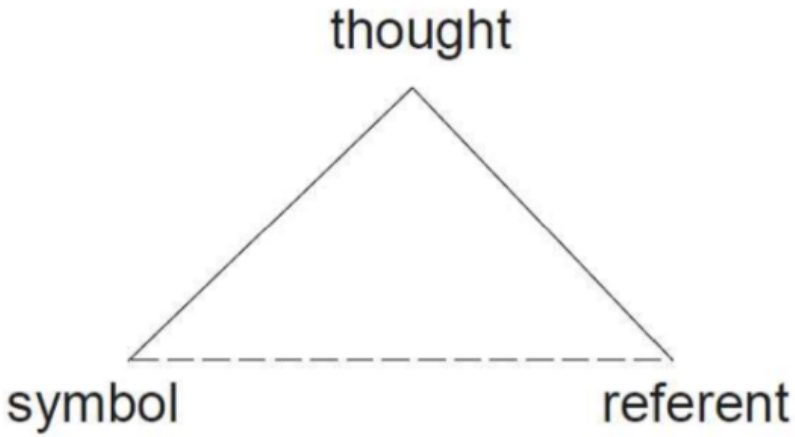
deskriptive Theorie:

- beschreibt, was der Fall ist
- Hauptinteresse der Linguistik

präskriptive Theorie:

- schreibt vor, was der Fall sein soll
- z.B. Rechtschreibreformen, nützlich beim Lernen einer Fremdsprache

Semiotisches Dreieck



Aspekte der Kommunikation mit sprachlichen Zeichen:

- symbol: Ausdrucksseite des sprachlichen Zeichens
  (das Wort “Baum”)
- thought: Inhaltsseite des sprachlichen Zeichens
  (das Konzept “Baum”, die Eigenschaften eines Baumes)
- referent: Gegenstand, Ereignis etc. in der außersprachlichen Wirklichkeit
  (Menge aller Bäume / ein bestimmter Baum)

Arbitrarität und Konventionalität

- Bedeutung B eines Ausdrucks A (der Ausdrucksseite eines Zeichens) ist im Allgemeinen nicht aufgrund von Eigenschaften von A vorhersagbar (vgl. z.B. Baum)
- In der Sprechergruppe hat sich die Konvention (Regel, Übereinkunft) durchgesetzt, A zu gebrauchen, wenn man B meint (vgl. z.B. Konvention, rechts zu fahren, nicht aber in England)
- Der Ausdruck A ist (in den meisten Fällen) willkürlich (arbiträr) der Bedeutung B zugeordnet

Arbitrarität und Konventionalität: Ausnahmen

- Ausnahme von der (völligen) Arbitrarität (aber nicht von der Konventionalität): Lautmalerei
- z.B. Bezeichnung für Gebell von Hunden wird in der Sprache nachgeahmt
- dt. wau wau (Kindersprache auch für Hund)
- engl. bow-wow
- russ. gav gav
- franz. ouah ouah
- Thai hoang hoang
- japan. kyankyan
- indones. gongong
- $\Rightarrow$ ist also nicht (bzw. nur sehr wenig) arbiträr, weil am realen Ereignis orientiert (Konvention ist aber dennoch vorhanden)

Übung

Welche Schwierigkeiten können bei der Distributionsanalyse auftreten, insbesondere in Schritt 2?
Distributionsanalyse: Verfahren zur Ermittlung sprachlicher Strukturen (amerikanischer Strukturalismus)

1. Segmentierung in Einheiten (Intuition, morphologische Anhaltspunkte)
2. Überprüfen der Segmente und zusammenfassen in paradigmatische Klassen anhand der Ersetzungsprobe.
3. Finden von Syntagmatischen Relationen zwischen den paradigmatischen Klassen.

Zum Schluss: Besonders klausurrelevant

- Wortform, Lexem, Token, Type
- Definitionsversuche des Wortkonzepts
  - Orthographisch, phonologisch, morphologisch, morphosyntaktisch, semantisch, intuitiv
- Paradigmatische vs. syntagmatische Sprachachse
- Distribution(sanalyse)
- Wohlgeformtheit, Deskriptivität vs. Präskriptivität
- Semiotisches Dreieck
- Arbitrarität