Einführung in die Computerlinguistik
Allgemeines zur Linguistik

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

2015-10-12

Take-away

- Was ist Linguistik / linguistische Grundbegriffe
- Das sprachliche Zeichen
- Das Wort

Overview

Outline

Linguistik (Sprachwissenschaft)

Theorie der natürlichen Sprachen

- d.h. natürliche Sprache beschreiben, modellieren
- natürliche Sprache: entstanden in einer historischen Entwicklung, z.B. Dt., Engl., Russ.
- nicht: Programmiersprachen, Logiksprachen, Plansprachen (z.B. Esperanto, Volapük)

Was ist Sprache?

- Man kann sich ein ganzes Studium lang nur mit dieser Frage beschäftigen.
- Teile dieser Vorlesung: Die Sicht einer Schule, der Moskauer Semantischen Schule
- System von Zeichen (Wortschatz) und Regeln (Grammatik) zur Mitteilung von Bedeutungen

Schulen in der Linguistik

- Strukturalismus
- Semiotik (zweiter Teil der Vorlesung)
- Wittgenstein
- Kognitive Linguistik
- Moskauer Semantische Schule (ist Grundlage eines Teils dieser Vorlesung)

Einige linguistische Grundbegriffe

- Token vs. Type
- Wortschatz/Vokabular
- Gebrauch vs. Erwähnung
- Meta- vs. Objektsprache
- Distribution
- Paradigmatische vs. syntagmatische Sprachachse
- Wohlgeformtheit und Grammatikalität
- Deskriptivität vs. Präskriptivität

Token vs. Type (1)

Token / Zeichenvorkommnis

- sinnlich wahrnehmbares, mündlich oder schriftlich geäußertes Zeichen
- konkrete Vorkommen eines Zeichens (Laut, Buchstabe, Wort, Satz, Text, . . .) an einem Ort zu einer Zeit

Type / Zeichentyp

- Klasse von Token, die von ihren Sprechern und Hörern (bzw. ihren Autoren und Lesern) nicht unterschieden werden und daher als gleich, z.B. als Kopien, wahrgenommen werden
- ursprüngliche Unterscheidung durch C.S. Peirce

“eine Rose ist eine Rose” $\rightarrow$ 5 Token, 3 Types

Token/Type: für Wörter oder Wortformen

- wichtig ist grundlegende Unterscheidung zwischen 'Wort' und 'Wortform' Beispiel: "eine Rose ist eine Rose und viele Rosen ergeben einen Strauß"
- Wortformen: $\rightarrow$ 11 Token, 9 Types
- Wörter: $\rightarrow$ 11 Token, 7 Types

Type/token ratio: Written language

http://www.speech-therapy-information-and-resources.com/downloads/type-token-ratio.pdf

TEXT 1: Written Language

But what are thoughts? Well, we all have them. They are variously described as ideas, notions, concepts, impressions, perceptions, views, beliefs, opinions, values, and so on. At times they are brief, coming and going in an instant. On other occasions they seem to endure and we can mull them over again and again in the act we call thinking. We can put them aside, fall asleep, and then return to them later. We refer to them as things we can handle. However, this is just a metaphor.

$\rightarrow$ 87 Token (Wortformen)

Type/token ratio: Written language (2)

rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	we	6	17	asleep	1	33	impressions	1	49	seem	1
2	and	5	18	at	1	34	instant	1	50	so	1
3	them	5	19	beliefs	1	35	is	1	51	the	1
4	are	3	20	brief	1	36	just	1	52	then	1
5	can	3	21	but	1	37	later	1	53	things	1
6	they	3	22	call	1	38	metaphor	1	54	thinking	1
7	to	3	23	coming	1	39	mull	1	55	this	1
8	again	2	24	concepts	1	40	notions	1	56	thoughts	1
9	as	2	25	described	1	41	occasions	1	57	times	1
10	in	2	26	endure	1	42	opinions	1	58	values	1
11	on	2	27	fall	1	43	other	1	59	variously	1
12	a	1	28	going	1	44	over	1	60	views	1
13	act	1	29	handle	1	45	perceptions	1	61	well	1
14	all	1	30	have	1	46	put	1	62	what	1
15	an	1	31	however	1	47	refer	1	**TOTAL**		**87**
16	aside	1	32	ideas	1	48	return	1			

$\rightarrow$ 62 Types

$$\text{Type-Token-Ratio} = \frac{\textit{AnzahlTypes}}{\textit{AnzahlToken}} = \frac{62}{87} = 0{,}713$$

Type/token ratio: Spoken language

TEXT 2: Speech

```
01      P:      so: (.) er: (..) as you were saying about er:: (.)
02              where are you living now Andrew
03      A:      Skipton Lodge
04      P:      Skipton Lodge?
05      A:      mm (...) Skipton Lodge
06      P:      yeah (.) do you like it
07      A:      yeah I do
08      P:      yeah
09      A:      I've settled in
10      P:      you have (...) good (.) w w what are the things you
11              like about it
12      A:      go out in the tow:n
13      P:      you go out in the town (...)
14      A:      yeah
15              (2.1)
16              with er: Tommy and Martin (.) and er:: (.) Noel
17      P:      and?
18      A:      NOEL
19      P:      oh yes (.) oh he lives there does he?
20      A:      yeah he live(s) in the flats
21      P:      yeah (.) oh they have flats there do they
22      A:      mm
23              (3.3)
24              and er::
25              (2.3)
26              and I went to see (..) (Elaine)
```

→

88 Token

Type/token ratio: Spoken language (2)

rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	yeah	6	13	flats	2	25	as	1	37	to	1
2	you	6	14	go	2	26	does	1	38	Tommy	1
3	and	5	15	have	2	27	Elaine	1	39	've	1
4	in	4	16	it	2	28	good	1	40	went	1
5	the	4	17	like	2	29	living	1	41	were	1
6	do	3	18	lives	2	30	Martin	1	42	what	1
7	he	3	19	Noel	2	31	now	1	43	where	1
8	I	3	20	out	2	32	saying	1	44	with	1
9	Lodge	3	21	there	2	33	see	1	45	yes	1
10	Skipton	3	22	they	2	34	settled	1			
11	about	2	23	town	2	35	so	1		**TOTAL**	**88**
12	are	2	24	Andrew	1	36	things	1			

$\rightarrow$ 45 Types

Type-Token-Ratio $= \frac{45}{88} = 0{,}511$

$\rightarrow$ etwa gleich viele Token wie in Text 1 $\rightarrow$ reichhaltigeres Vokabular in Text 1

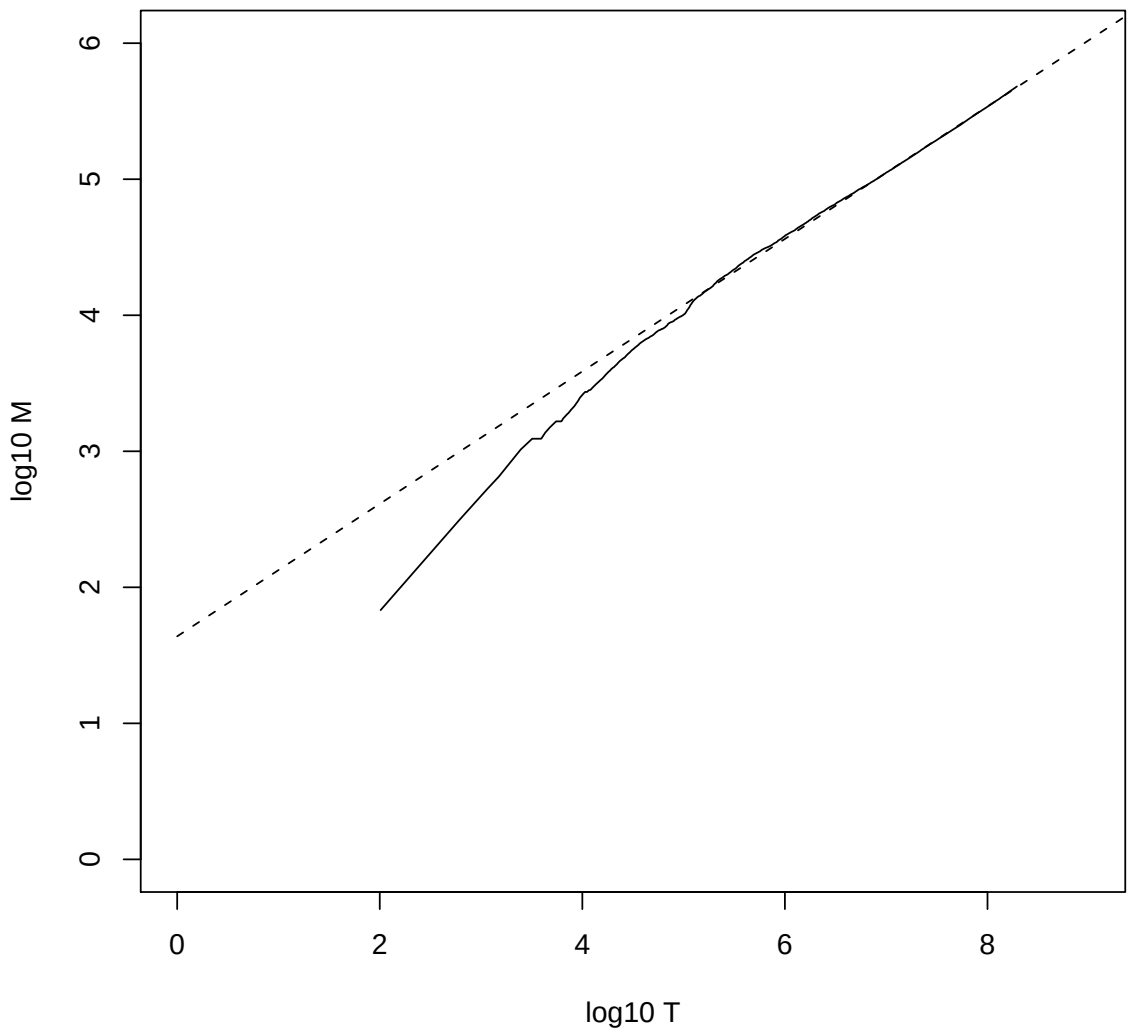
Abdeckung des Wortschatzes durch ein Korpus

- Type-Token-Ratio ist Indiz der Abdeckung eines Korpus (in Bezug auf Wörter/Wortformen)
- Typischerweise flacht diese Funktion (x-Achse: Token, y-Achse: Types) nach steiler Anfangskurve asymptotisch ab
- wenn dies eintritt, dann heißt das, dass eine weitere Vergrößerung des Korpus die Abdeckung des Wortschatzes nur geringfügig verbessern wird.

How big is the term vocabulary?

- That is, how many types (distinct words) are there?
- Can we assume there is an upper bound?
- Not really: At least $70^{20} \approx 10^{37}$ different words of length 20.
- The vocabulary will keep growing with collection size.
- Heaps' law: $M = kT^b$
- M is the size of the vocabulary, T is the number of tokens in the collection.
- Typical values for the parameters k and b are: $30 \leq k \leq 100$ and $b \approx 0.5$.
- Heaps' law is linear in log-log space.
  - It is the simplest possible relationship between collection size and vocabulary size in log-log space.
  - Empirical law

Heaps' law for Reuters



Vocabulary size M as a function of collection size T (number of tokens) for Reuters-RCV1. For these data, the dashed line $\log_{10} M = 0.49 * \log_{10} T + 1.64$ is the best least squares fit. Thus, $M = 10^{1.64} T^{0.49}$ and $k = 10^{1.64} \approx 44$ and $b = 0.49$.

Empirical fit for Reuters

- Good, as we just saw in the graph.
- Example: for the first 1,000,020 tokens Heaps' law predicts 38,323 terms:

$$44 \times 1{,}000{,}020^{0.49} \approx 38{,}323$$

- The actual number is 38,365 terms, very close to the prediction.
- Empirical observation: fit is good in general.

Exercise

1. What is the effect of including spelling errors vs. automatically correcting spelling errors on Heaps' law?
2. Compute vocabulary size M
   - Looking at a collection of web pages, you find that there are 3000 different terms in the first 10,000 tokens and 30,000 different terms in the first 1,000,000 tokens.
   - Assume a search engine indexes a total of 20,000,000,000 (2×10^{10}) pages, containing 200 tokens on average
   - What is the size of the vocabulary of the indexed collection as predicted by Heaps' law?

Gebrauch vs. Erwähnung

Gebrauch (*engl.* use):

- Ein Wort (oder ein anderes sprachliches Zeichen), mit dessen Hilfe über etwas gesprochen wird (also über dessen Denotat), heißt "gebraucht"
- z.B. bzgl. "Hans":

  (3) Es ist fraglich, ob Hans ein guter Vater für das Kind ist.

Erwähnung (Anführung, *engl.* mention):

- Ein Wort, über das gesprochen wird, heißt "erwähnt"

  (4) Es ist fraglich, ob "Hans" ein guter Name für das Kind ist.

Meta- vs. Objektsprache

- Metasprache: Sprache, in der man spricht bzw. etwas beschreibt
- Objektsprache: Sprache, über die bzw. über deren Ausdrücke (Zeichen) man spricht
- z.B. auf deutsch über die englische Grammatik sprechen

Distribution

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”

Paradigmatische vs. syntagmatische Sprachachse

paradigmatische Sprachachse:

- Beziehung von einem Zeichen (Wörtern, Wortformen) zu anderen Zeichen des gleichen Paradigmas
- Ebene der Ersetzung

syntagmatische Sprachachse:

- Beziehung von einem Zeichen (Wortformen) zu Zeichen in seinem Kontext (in einem konkreten Satz)
- Ebene der Kombination

Beispiel zu paradigmatische vs. syntagmatische Sprachachse

(5)

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

← syntagmatisch →

Wie könnte man dieses Schema erweitern?

- paradigmatisch: *Natascha steht am Fenster*
- syntagmatisch: *. . . und denkt über Linguistik nach*

Wohlgeformtheit / Grammatikalität

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
- Noam Chomsky (1957):

  (6) Colorless green ideas sleep furiously. vs. *Ideas green sleep colorless furiously
- nicht wohlgeformte Sätze (Ausdrücke) werden mit Stern gekennzeichnet

Deskriptivität vs. Präskriptivität

deskriptive Theorie:

- beschreibt, was der Fall ist

präskriptive Theorie:

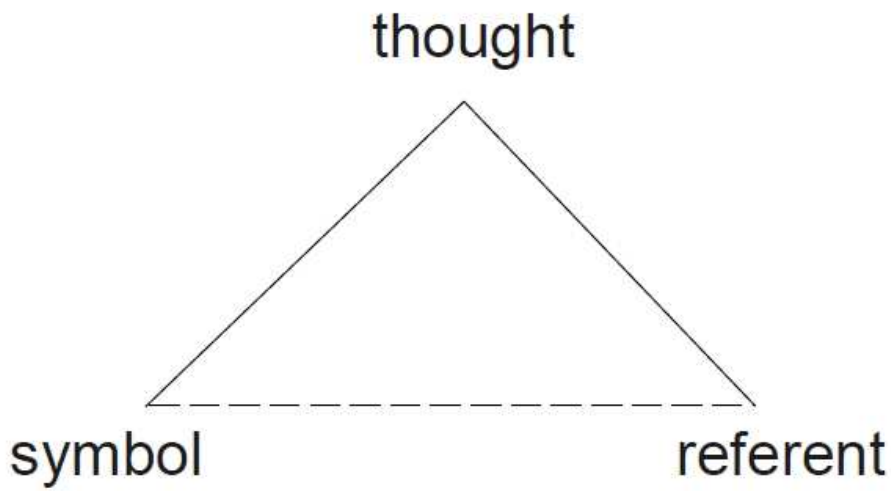
- schreibt vor, was der Fall sein soll

Literatur: Bußmann, H.: Lexikon der Sprachwissenschaft. Stuttgart 2008 Lewandowski, Th.: Linguistisches Wörterbuch. Heidelberg 1994

Outline

Semiotisches Dreieck

Ogden, Richards: The Meaning of Meaning (1923)



Alternative Bezeichnungen im semiotischen Dreieck

symbol:

- signifiant (Saussure)
- Signifikant
- Bezeichnendes

$\rightarrow$ Ausdrucksseite des sprachlichen Zeichens

Alternative Bezeichnungen im semiotischen Dreieck

thought:

- signifié (Saussure)
- Signifikat (Morris)
- Bezeichnetes
- Sinn (bei Frege)
- Bedeutung (außer bei Frege)
- Intension (Carnap)

$\rightarrow$ Inhaltsseite des sprachlichen Zeichens

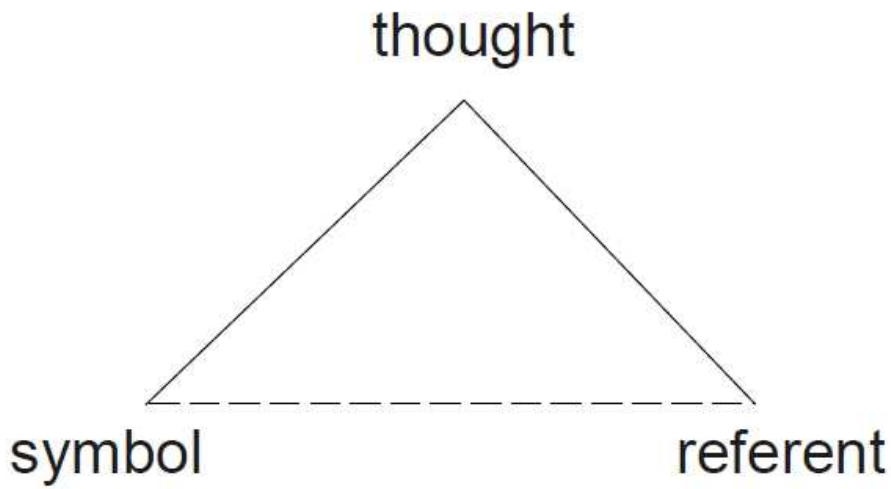
Alternative Bezeichnungen im semiotischen Dreieck

referent:

- Denotat (Morris)
- Extension (Carnap)
- Bedeutung (bei Frege)

$\rightarrow$ etwas, das seinen Platz in der außersprachlichen Wirklichkeit hat: Gegenstand oder Ereignis etc., „Ding“

Semiotisches Dreieck



- symbol: Ausdrucksseite des sprachlichen Zeichens
- thought: Inhaltsseite des sprachlichen Zeichens
- referent: Gegenstand, Ereignis etc. in der außersprachlichen Wirklichkeit

Arbitrarität und Konventionalität natürlichsprachlicher Zeichen (1)

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

Aufgabe:
Zeichen, die nicht (vollständig) arbiträr sind

- Bedeutung B eines Zeichens Z (genauer: Ausdrucksseite) ist im Allgemeinen nicht aufgrund von Eigenschaften von Z vorhersagbar (vgl. z.B. *Tisch*)
- In der Sprechergruppe hat sich die Konvention (Regel, Übereinkunft) durchgesetzt, Z zu gebrauchen, wenn man B meint (vgl. z.B. Konvention, rechts zu fahren, nicht aber in England)
- Zeichen Z ist (in den meisten Fällen) willkürlich (arbiträr) der Bedeutung B zugeordnet
- Finden Sie Zeichen Z, die Ausnahmen von dieser Generalisierung sind.

Arbitrarität und Konventionalität natürlichsprachlicher Zeichen (2)

- Ausnahme von der (völligen) Arbitrarität (aber nicht von der Konventionalität) ist Onomatopöie (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)

Outline

Übung

Wie viele Wörter gibt es jeweils in folgenden Sätzen:

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.

Criteria for wordhood

- orthographic / graphemic
- phonological
- morphological
- morphosyntactic
- semantic
- “intuition”
- Literature: Heringer, H.-J.: Morphologie. Paderborn 2009

Orthographische/graphematische Kriterien

"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!

Phonologische Kriterien

"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”

b) “Wörter sind solche flektierbaren grammatische Einheiten, die über eine einheitliche Flexion verfügen.” aber:

- nicht flektierbare Wörter?!

Morphosyntaktische Kriterien

"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*
- mehrere Wörter für einen Begriff! z.B. *roter Faden*, *Frankfurter-Straßennamen-Büchlein*, engl. *jam sandwich*

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’.

What is a word: Summary

- In many cultures, there is a clear, intuitive notion of what a word is.
- Based on orthographic / graphemic, phonological, morphological, morphosyntactic, semantic criteria
- The intuitive notion of "word" is not a clearcut concept.
- Rather it is a prototype / family resemblance concept – some words violate some of the criteria.
- There are individual cases where people have different / unclear intuitions: "Rad fahren" vs "radfahren".
- Different cultures have different notions of word: "business trip" is two words in English, "Dienstreise" is one word in German (although they are very similar linguistically).
- Certain linguistic theories aim to come up with precise definitions of wordhood ...
- ... but these can deviate significantly from the intuitive notion of what a word is.

Take-away

- Was ist Linguistik / linguistische Grundbegriffe
- Das sprachliche Zeichen
- Das Wort