Introduction

VL Sprachliche Informationsverarbeitung

Nils Reiter

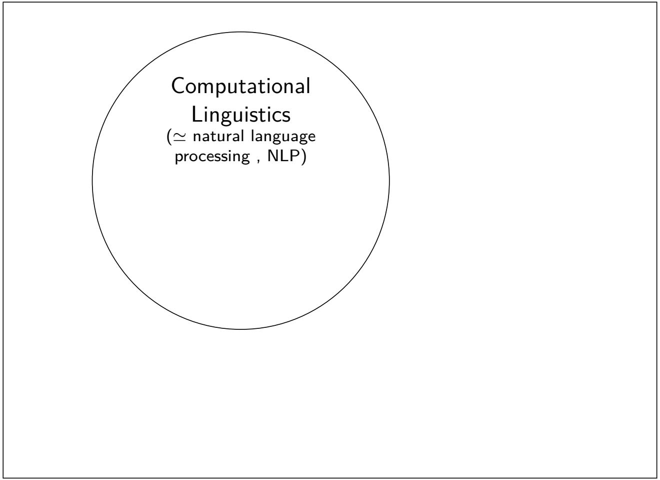
nils.reiter@uni-koeln.de

October 13, 2022

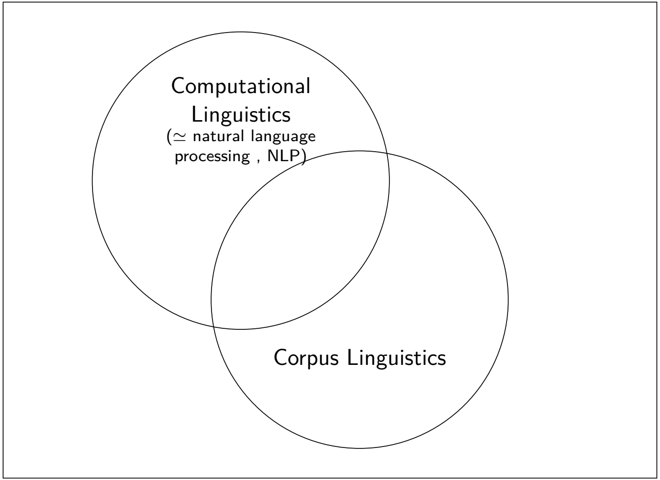
Winter term 2022/23

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

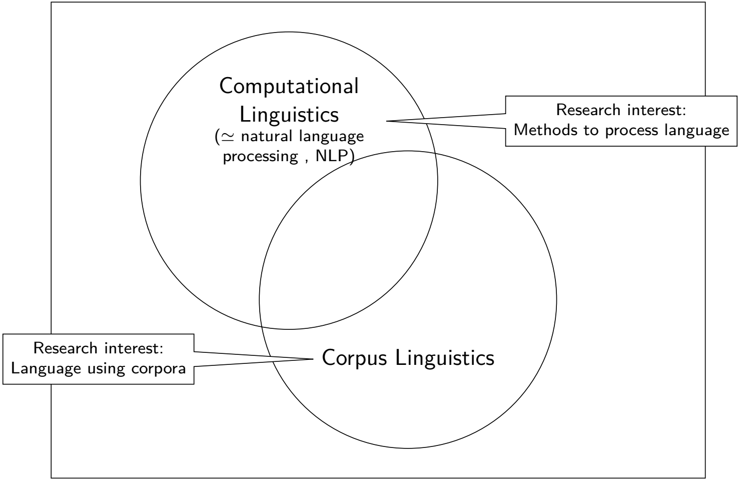
Disciplinary Placement


Computational Linguistics
(≃ natural language processing , NLP)

Disciplinary Placement

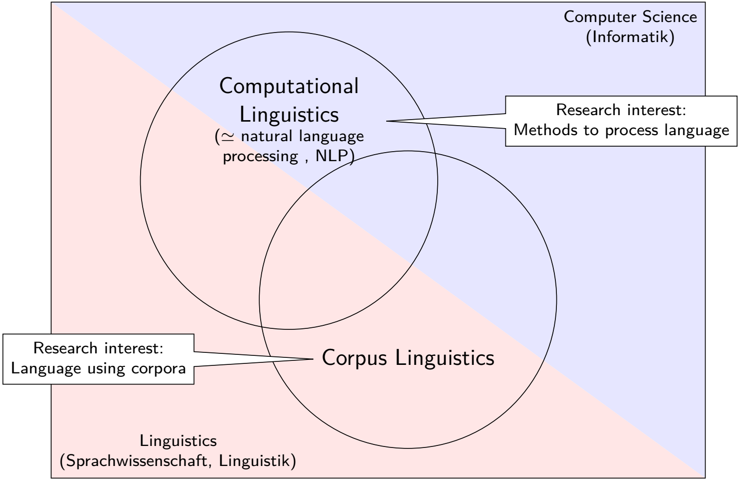

Computational Linguistics
(≃ natural language processing , NLP)
Corpus Linguistics

Disciplinary Placement



Computational Linguistics
(≃ natural language processing , NLP)
Research interest:
Methods to process language
Research interest:
Language using corpora
Corpus Linguistics

Disciplinary Placement



Brief history of Computational Linguistics I

- 1950s: DARPA Projects to automatically translate Russian into English
- 1957/65: Linguistics shifts focus from describing to generating Chomsky (1957, 1965)
- 1959: Theo Lutz for the first time generates a German poem with a computer Lutz (1959)
- 1962: Foundation of the „Association for Machine Translation and Computational Linguistics“, 1968 renamed to „Association for Computational Linguistics (ACL)“
- 1966, ALPAC report: MT more expensive, less accurate and slower than human translation ALPAC (1966)
- 1968: Foundation of SYSTRAN, first MT company
- 1975: European commission uses SYSTRAN software (first use of MT on EU level)

Brief history of Computational Linguistics II

- 1984: First corpus-based commercial MT system Nagao (1984)
- 1992: Study programs established in Germany (Saarbrücken/Stuttgart)
- 2011: IBM Watson beats two humans in Jeopardy YouTube / Apples Siri launched
- 2013: Word embeddings (e.g., word2vec) Mikolov u. a. (2013)
- 2017: Launch of the DeepL Translator
- 2018: Transformer models: BERT Devlin u. a. (2019)

Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ‚digital methods' on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, 4)
  - … but this is a minority position in CL, often also seen as part of AI

Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ‚digital methods‘ on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, 4)
  - … but this is a minority position in CL, often also seen as part of AI
  - Historically (and still today) split between engineering (natural language processing, NLP) and science/scholarship (computational linguistics, CL)
  - ⚠ Neurolinguistic programming and natural language processing are not the same (both use ‚NLP‘ as abbreviation)

Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ‚digital methods‘ on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, 4)
  - … but this is a minority position in CL, often also seen as part of AI
  - Historically (and still today) split between engineering (natural language processing, NLP) and science/scholarship (computational linguistics, CL)
  - ⚠ Neurolinguistic programming and natural language processing are not the same (both use ‚NLP‘ as abbreviation)

University of Cologne

For historic reasons, CL and NLP are called „Sprachliche Informationsverarbeitung“

Institut für Digital Humanities

Historisch-Kulturwissenschaftliche Informationsverarbeitung

- Prof. Dr. Øyvind Eide
- Keywords
  - Maps
  - Models and modeling
  - Cultural heritage
  - Simulation

Sprachliche Informationsverarbeitung

- Prof. Dr. Nils Reiter
- Keywords
  - Geschriebene und gesprochene Sprache
  - Textanalyse
  - Machine/deep learning
  - Operationalisierung

Experiments

- Cornerstone of the ‚scientific method‘
- Used in many disciplines: Natural sciences, social sciences, medicine, …

Experiments

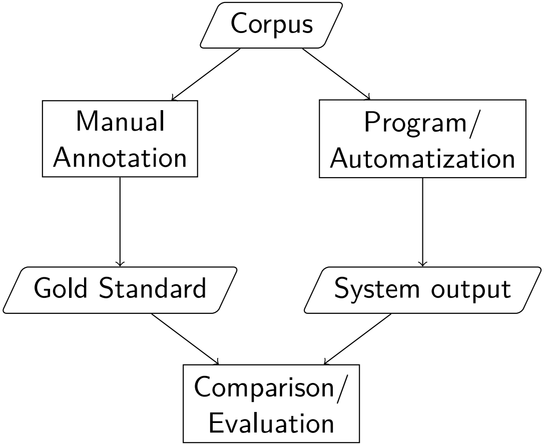
- Cornerstone of the ‚scientific method‘
- Used in many disciplines: Natural sciences, social sciences, medicine, …
- Experiments are used to verify or falsify hypotheses
- Reproducibility: The outcome does not depend on the experimenter

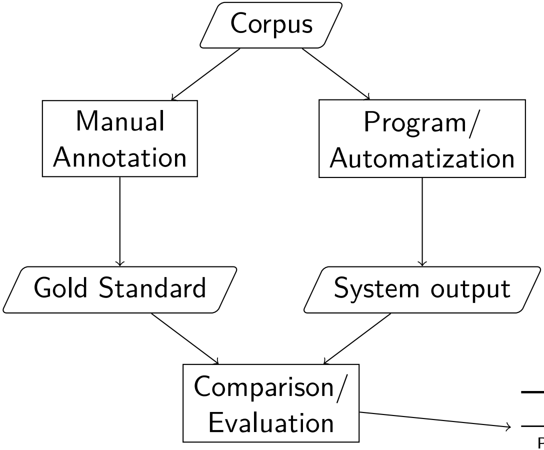
Experiments

- Cornerstone of the ‚scientific method'
- Used in many disciplines: Natural sciences, social sciences, medicine, …
- Experiments are used to verify or falsify hypotheses
- Reproducibility: The outcome does not depend on the experimenter
- CL: Hypotheses about the operationalisation of language/text phenomena

Example

Position within a sentence is indicative for the part of speech


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation



Corpus

Manual Annotation

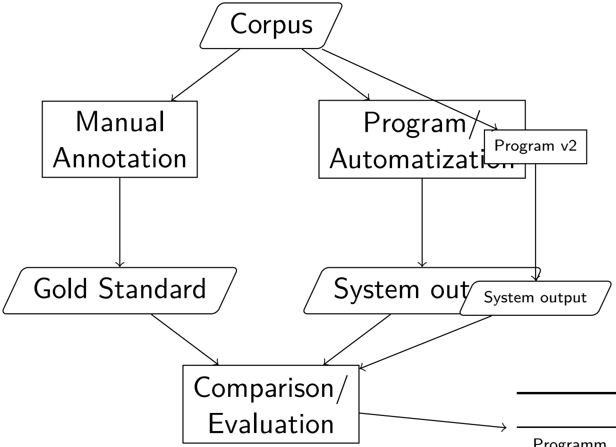
Program/ Automatization

Gold Standard

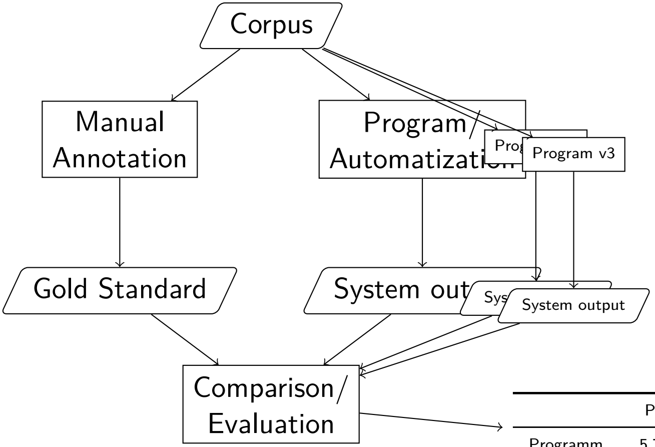
System output

Comparison/ Evaluation

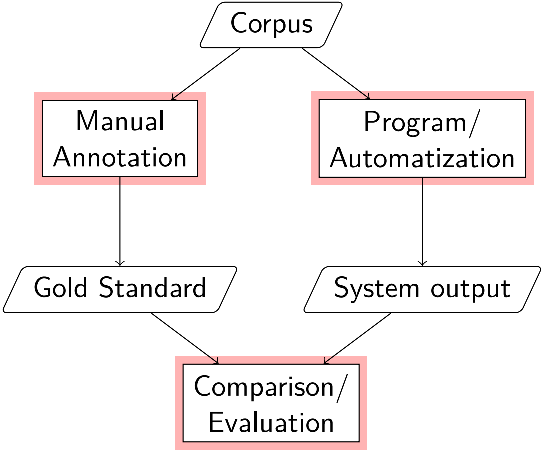
	P	R	F
Programm	5.7	9.2	...



	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...



	P	R	F
Programm	5.7	9.2	…
v2	9.9	16.7	…
v3	15.3	21.8	…


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation

Section 2

Organisatorisches

Introduction

Organisatorisches

Language and Linguistics

- Phonology and Phonetics
- Morphology
- Syntax

Orga

- Donnerstag, 12:00-13:30
- Module: ...
- Studienleistung: fünf Hausaufgaben, Abgabe via Ilias
- Prüfung: Klausur (02.02.2023)

Kurswebseite

Section 3

Language and Linguistics

Introduction

Organisatorisches

Language and Linguistics

- Phonology and Phonetics
- Morphology
- Syntax

What is Linguistics?

> *Linguistics is the scientific study of language.*
>
> Wikipedia, 925699120

What is Linguistics?

> *Linguistics is the scientific study of language.*
>
> Wikipedia, 925699120

- ‚Scientific study'
  - ‚the' scientific method
  - Testable explanations
- Language
  - ?

What is Linguistics?

> *Linguistics is the scientific study of language.*
>
> Wikipedia, 925699120

- ‚Scientific study'
  - ‚the' scientific method
  - Testable explanations
- Language
  - ?

Prescriptive vs. descriptive

- Prescriptive: Telling people how to use language
- Descriptive: Observing and analysing how people do use language

What is Linguistics?

> *Linguistics is the scientific study of language.*
>
> Wikipedia, 925699120

- ‚Scientific study‘
  - ‚the‘ scientific method
  - Testable explanations
- Language
  - ?

Prescriptive vs. descriptive

- Prescriptive: Telling people how to use language
- Descriptive: Observing and analysing how people do use language
- Academic linguistics: Nowadays mostly descriptive

What is Language?

- Communication system
- Conventionalised: We agree (mostly)
  - Only partially authoritative

What is Language?

- Communication system
- Conventionalised: We agree (mostly)
  - Only partially authoritative
- What do we agree on?
  - Relation between *sign*s and its *meaning* (which is not the same!)
    - Saussure: Semiotics
  - E.g.: ‚the students in this class‘ *means* all of you

Linguistic sign

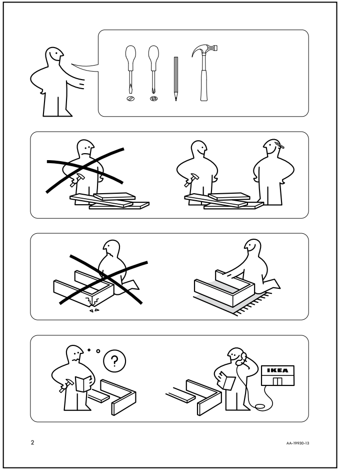
- Abstract notion
- Different levels
  - Texts
  - Sentences
  - Words
  - Syllables, morphemes
  - Spoken utterances

Linguistic sign

- Abstract notion
- Different levels
  - Texts
  - Sentences
  - Words
  - Syllables, morphemes
  - Spoken utterances
  - Non-textual signs
    - Emojis 😍

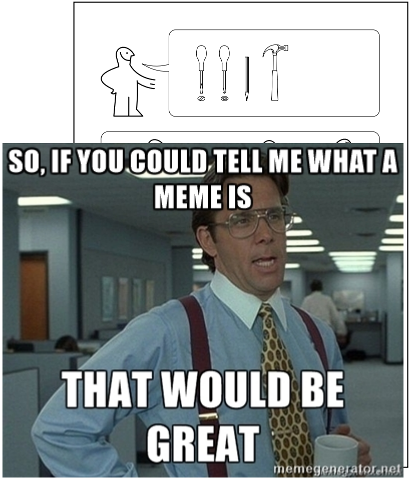
Linguistic sign

- Abstract notion
- Different levels
  - Texts
  - Sentences
  - Words
  - Syllables, morphemes
  - Spoken utterances
  - Non-textual signs
    - Emojis 😍
    - Assembly instructions



Linguistic sign

- Abstract notion
- Different levels
  - Texts
  - Sentences
  - Words
  - Syllables, morphemes
  - Spoken utterances
  - Non-textual signs
    - Emojis 😍
    - Assembly instructions
    - Memes
    - ...



Linguistic sign

Meaning is arbitrary

- ‚dog‘ refers to a four-legged, wolf-like mammal

Linguistic sign

Meaning is arbitrary

- ‚dog' refers to a four-legged, wolf-like mammal
- This is an arbitrary decision
- The sign could be different, e.g., ‚cat'
- No inherent meaning in signs
  - ...but strongly conventionalised

Linguistic sign

Meaning is arbitrary

- ‚dog‘ refers to a four-legged, wolf-like mammal
- This is an arbitrary decision
- The sign could be different, e.g., ‚cat‘
- No inherent meaning in signs
  - …but strongly conventionalised
- Interpreting signs (and language) is something we learn
- Language is a social construct
  - Studying language is different from studying gravity

Ambiguities

- Der Jäger traf den Mann mit dem Gewehr.

Ambiguities

- Der Jäger traf den Mann mit dem Gewehr.
- Landesmusikdirektor:in

Ambiguities

- Der Jäger traf den Mann mit dem Gewehr.
- Landesmusikdirektor:in
- Maria hat Petra beim Einkaufen getroffen. Sie hat ihr Geld geliehen.

Ambiguities

- Der Jäger traf den Mann mit dem Gewehr.
- Landesmusikdirektor:in
- Maria hat Petra beim Einkaufen getroffen. Sie hat ihr Geld geliehen.
- hubert hat dort liebe genossen.
- ...

Ambiguities

- Der Jäger traf den Mann mit dem Gewehr.
- Landesmusikdirektor:in
- Maria hat Petra beim Einkaufen getroffen. Sie hat ihr Geld geliehen.
- hubert hat dort liebe genossen.
- ...

Linguistics: Let's explain / represent / reproduce these ambiguities

Introduction

Organisatorisches

Language and Linguistics

Phonology and Phonetics

Morphology

Syntax

Phonology and Phonetics

Phonetics

- How are language sounds produced and understood/processed?
- Focus: Practical, verbal and gestural use of language
- Links to biology, acoustics

Phonology

- Which function have certain phonemes within a language?
- Focus: Relation to other areas of linguistics and grammar
- Abstraction over concrete phonemes

Understanding Spoken Language

Relevant and irrelevant differences

- [ʃaːl] vs. [ʃal] (Schal vs. Schall)
  - Vowel length indicates a difference in meaning

Understanding Spoken Language

Relevant and irrelevant differences

- [ʃaːl] vs. [ʃal] (Schal vs. Schall)
  - Vowel length indicates a difference in meaning
- [roːt] vs. [ʀoːt] (rot)
  - Pronunciation of /r/ doesn't make a difference (in German)

Understanding Spoken Language

Relevant and irrelevant differences

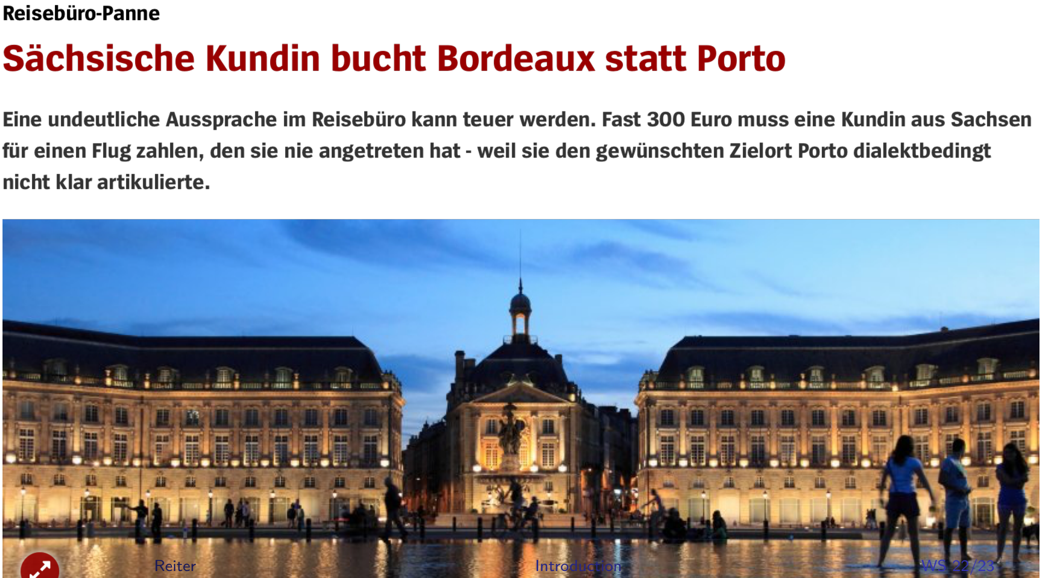
- [ʃaːl] vs. [ʃal] (Schal vs. Schall)
  - Vowel length indicates a difference in meaning
- [roːt] vs. [ʀoːt] (rot)
  - Pronunciation of /r/ doesn't make a difference (in German)

International Phonetic Alphabet (IPA)

https://www.internationalphoneticassociation.org

- Symbols defined via physiological properties of the pronounciation

Pronunciation mishaps



Reisebüro-Panne

Sächsische Kundin bucht Bordeaux statt Porto

Eine undeutliche Aussprache im Reisebüro kann teuer werden. Fast 300 Euro muss eine Kundin aus Sachsen für einen Flug zahlen, den sie nie angetreten hat - weil sie den gewünschten Zielort Porto dialektbedingt nicht klar artikulierte.

Pronunciation mishaps

Bordeaux vs. Porto

- Porto: [ˈpɔʁto]
- Bordeaux: [bɔʁˈdoː]

Pronunciation mishaps

Bordeaux vs. Porto

- Porto: [ˈpɔʁto]
- Bordeaux: [bɔʁˈdoː]
- Key difference: Voicing of the plosives p/b and t/d
  - /p/, /t/: voiceless (stimmlos)
  - /b/, /d/: voiced (stimmhaft)

Pronunciation mishaps

Bordeaux vs. Porto

- Porto: [ˈpɔʁto]
- Bordeaux: [bɔʁˈdoː]
- Key difference: Voicing of the plosives p/b and t/d
  - /p/, /t/: voiceless (stimmlos)
  - /b/, /d/: voiced (stimmhaft)

Voice and Plosives

- Voice
  - Sounds with the use of the larynx (dt. Stimmlippen)
  - Example: Phase (voiceless: /f/) vs. Vase (voiced: /v/)
  - You can feel voice if you touch your throat

Pronunciation mishaps

Bordeaux vs. Porto

- Porto: [ˈpɔʁto]
- Bordeaux: [bɔʁˈdoː]
- Key difference: Voicing of the plosives p/b and t/d
  - /p/, /t/: voiceless (stimmlos)
  - /b/, /d/: voiced (stimmhaft)

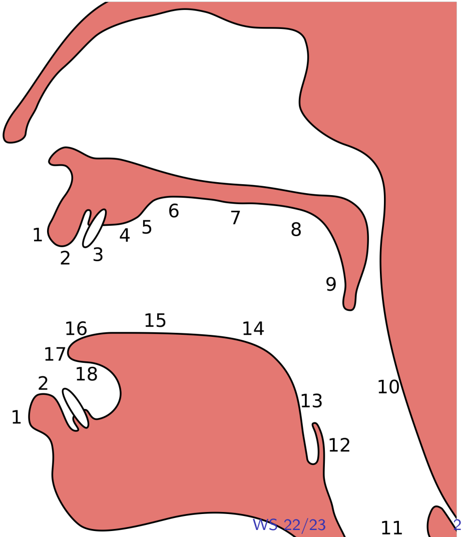
Voice and Plosives

- Voice
  - Sounds with the use of the larynx (dt. Stimmlippen)
  - Example: Phase (voiceless: /f/) vs. Vase (voiced: /v/)
  - You can feel voice if you touch your throat
- Plosive
  - Air stream is blocked, but suddenly re-opened
  - Example: /bʊs/ (plosive) vs. /mʊs/ (nasal)

Producing Sounds

Important Locations for German Sounds (Consonants)

2. labial (Lippen): [b], [p]
3. dental (Zähne): [v], [f]
4. alveolar (Zahnfach): [d], [t], ...
5. postalveolar: [ʃ]
7. palatal: [ç]
8. velar: [g], [k], ...
11. glottal: [ʔ]
    - ‚ein Echo‘: [aɪn ʔɛço]
    - ‚Student:in‘: [ʃtuˈdɛntʔɪn]



Producing Sounds

Consonants vs. Vowels

- Consonant
  - Produced with (complete or partial) closure of the vocal tract
  - labial/dental/... describes the position of the closure in the tract

Producing Sounds

Consonants vs. Vowels

- Consonant
    - Produced with (complete or partial) closure of the vocal tract
    - labial/dental/... describes the position of the closure in the tract
- Vowel
    - Produced without closure of the vocal tract
    - Usually voiced
    - Shaped by tongue position and lip rounding
        - (this is a simplification)

Subsection 2

Morphology

Introduction

Organisatorisches

Language and Linguistics

- Phonology and Phonetics
- Morphology
- Syntax

Morphology

- How do we create words?

Morphology

- How do we create words?
- Ambiguity:
  - Order in which parts of words are assembled

Morphology

- How do we create words?
- Ambiguity:
  - Order in which parts of words are assembled
- Morphological processes are language-dependent
  - German: Nominal composition
    - Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

Subsection 3

Syntax

Introduction

Organisatorisches

Language and Linguistics

- Phonology and Phonetics
- Morphology
- Syntax

Syntax

- Syntax: How are words used to form sentences?
  - Related to ‚grammar'
  - Two ways to look at syntax
    - Phrase structure
    - Dependency (not today)

Phrase Structure

- Words are not put in any arbitrary order
- Parts of speech (Wortarten) are not enough to explain sentences

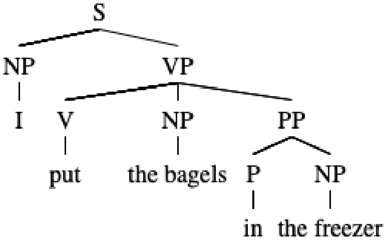
Phrase Structure

- Words are not put in any arbitrary order
- Parts of speech (Wortarten) are not enough to explain sentences
- Constituents
  - Words that are grouped together as a unit
  - What can appear in diff. positions of a sentence is a constituent

(1) I put the bagels in the freezer.

(2) The bagels, I put in the freezer.

(3) I put in the fridge the bagels (that John had given me).



Phrase Structure

Heads

- Phrases have heads
- Heads determine syntactic properties of the phrase
    - E.g., if the head is in plural, the phrase is in plural

Phrase Structure

Heads

- Phrases have heads
- Heads determine syntactic properties of the phrase
  - E.g., if the head is in plural, the phrase is in plural
- Dependent elements follow the head
  - Agreement

Phrase Structure

Heads

- Phrases have heads
- Heads determine syntactic properties of the phrase
  - E.g., if the head is in plural, the phrase is in plural
- Dependent elements follow the head
  - Agreement

Examples

(1) $\overbrace{\text{Der} \quad \text{flauschige} \quad \underbrace{\text{Hund}}_{\text{Head}}}^{\text{Nominal phrase}} \quad \text{bellt} \quad .$

(2) $\overbrace{\text{Die} \quad \text{flauschigen} \quad \underbrace{\text{Hunde}}_{\text{Head}}}^{\text{Nominal phrase}} \quad \text{bellen} \quad .$

German Syntax

Peculiarities in German (*every* language has their share of oddities)

German Syntax

Peculiarities in German (*every* language has their share of oddities)

- Free word order
  - ‚Den Hund hat er gestreichelt.'
  - ‚Er hat den Hund gestreichelt.'

German Syntax

Peculiarities in German (*every* language has their share of oddities)

- Free word order
  - ‚Den Hund hat er gestreichelt.‘
  - ‚Er hat den Hund gestreichelt.‘
- Separable verbs

German Syntax

Peculiarities in German (*every* language has their share of oddities)

- Free word order
    - ‚Den Hund hat er gestreichelt.‘
    - ‚Er hat den Hund gestreichelt.‘
- Separable verbs
    - aufstehen: ‚Sie steht jeden Tag früh auf.‘
        - *‚Sie aufsteht jeden Tag früh‘
    - bestehen: ‚Sie besteht die Prüfung.‘
        - *‚Sie steht die Prüfung be.‘
    - Mark Twain: ‘The Germans have another kind of parenthesis, which they make by splitting a verb in two and putting half of it at the beginning of an exciting chapter and the other half at the end of it. Can any one conceive of anything more confusing than that?’

German Syntax

Nominal Phrases

NP → Artikel? Adjektiv* Nomen (PP|Relativsatz)*

? 0 oder 1 mal
* 0 mal oder öfter
(|) Alternative