Morphology

The study of the way words are built up from smaller meaning-bearing units

Morphemes

- A *morpheme* is the smallest meaning-bearing unit of a language
- A *stem* is the central morpheme of the word, supplying the main meaning
- Affixes: Bits and pieces that adhere the stems (often with grammatical functions)

Word Formation

- **Words arise**
- **A new word „unhappy“ can be derived by left-concatenation of the prefix „un“ to the word „happy“**
- **„unhappy“ and „happy“ are two different words**

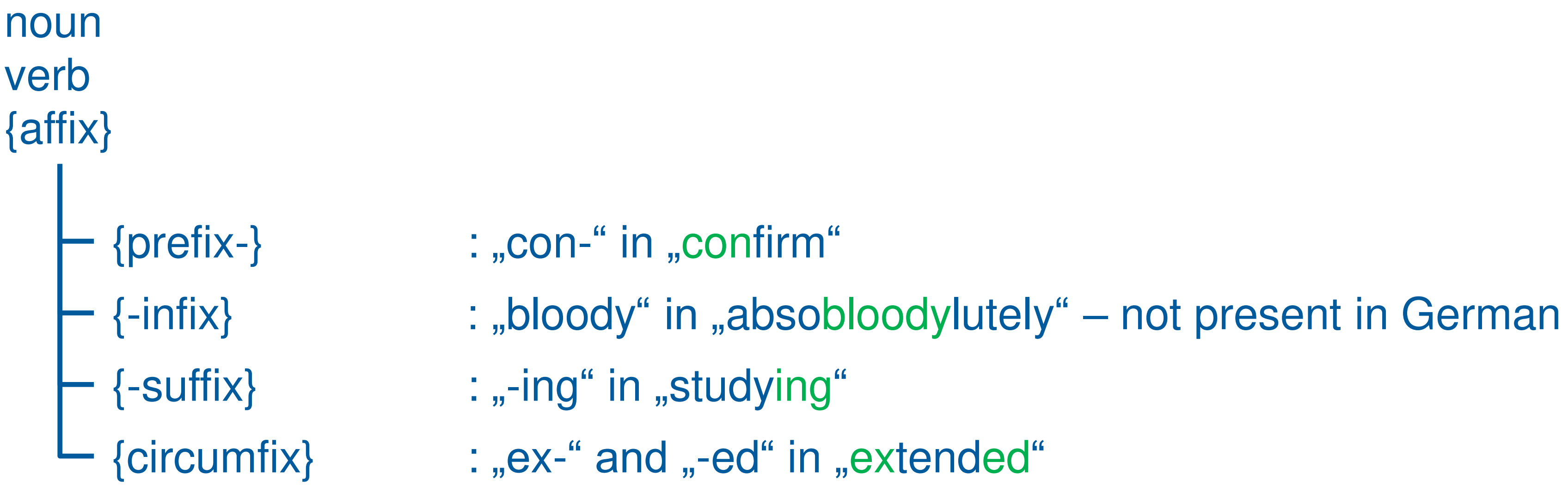
https://de.wikipedia.org/wiki/Wortbildung

Inflection

- Expresses grammatical functions of words in the sentence
- We can create the word „cats“ via inflection of the word „cat“ using the plural „-s“
- „cat“ and „cats“ are two forms of the same word

https://de.wikipedia.org/wiki/Flexion

Morphology



noun
verb
{affix}

{prefix-}	: „con-“ in „confirm“
{-infix}	: „bloody“ in „absobloodylutely“ – not present in German
{-suffix}	: „-ing“ in „studying“
{circumfix}	: „ex-“ and „-ed“ in „extended“

Interfix, duplifix, transflix, simulfix, supraflix, disfix, …

https://en.wikipedia.org/wiki/Affix
https://en.wiktionary.org/wiki/absobloodylutely

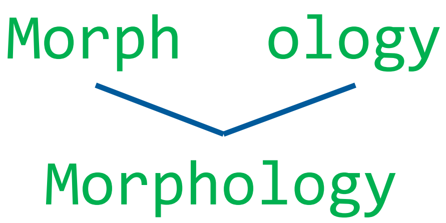
Morphology Tree: Example 1

Morphology

unbeliefable

https://www.youtube.com/watch?v=QT_A-7usieI&feature=emb_rel_end
https://all-about-linguistics.group.shef.ac.uk/branches-of-linguistics/morphology/what-is-morphology/

Morphology Tree: Example 1

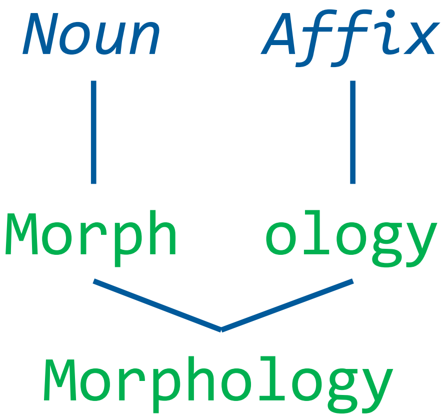


Morph ology

Morphology

unbeliefable

Morphology Tree: Example 1



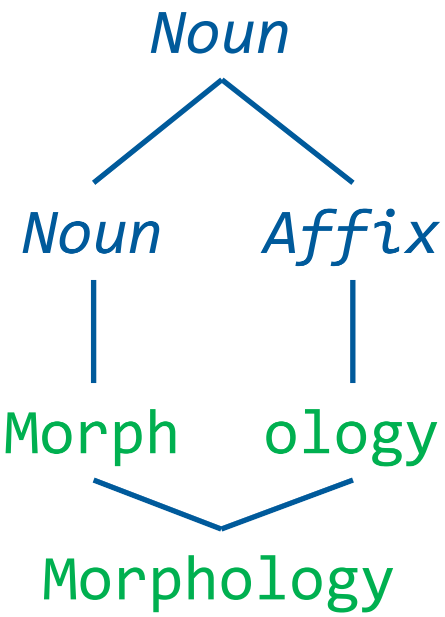
Noun *Affix*

Morph ology

Morphology

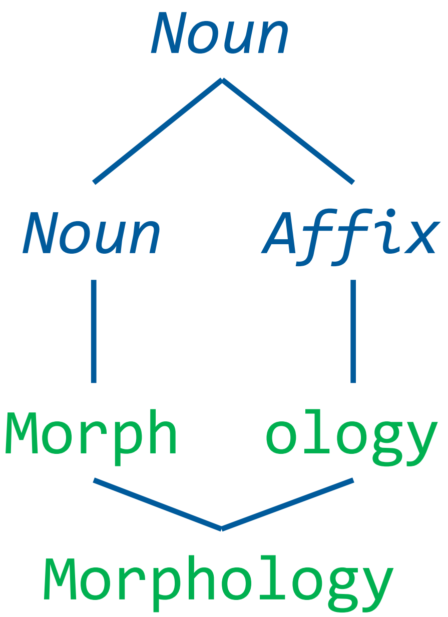
unbeliefable

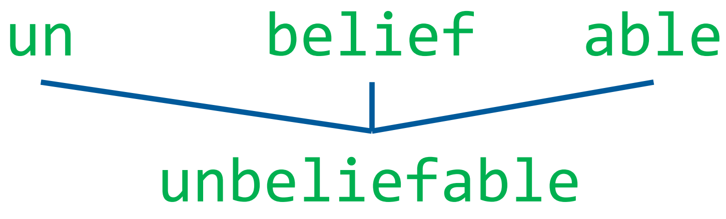
Morphology Tree: Example 1



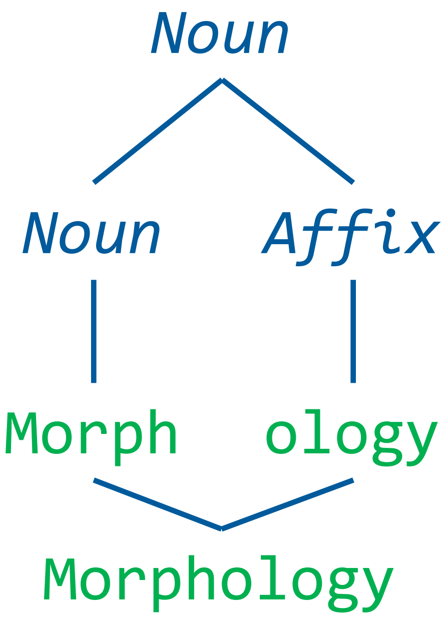
unbeliefable

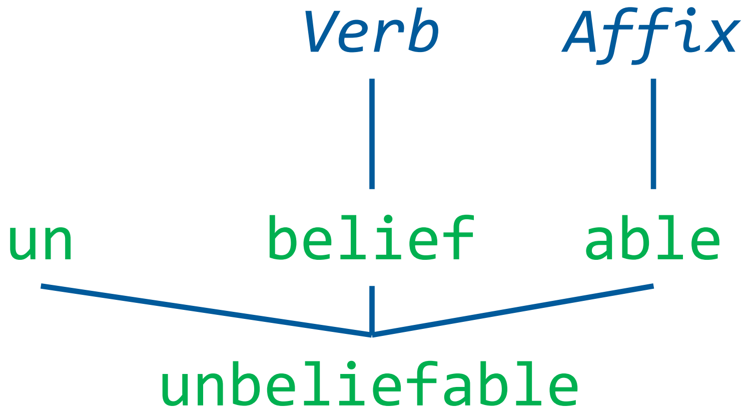
Morphology Tree: Example 1



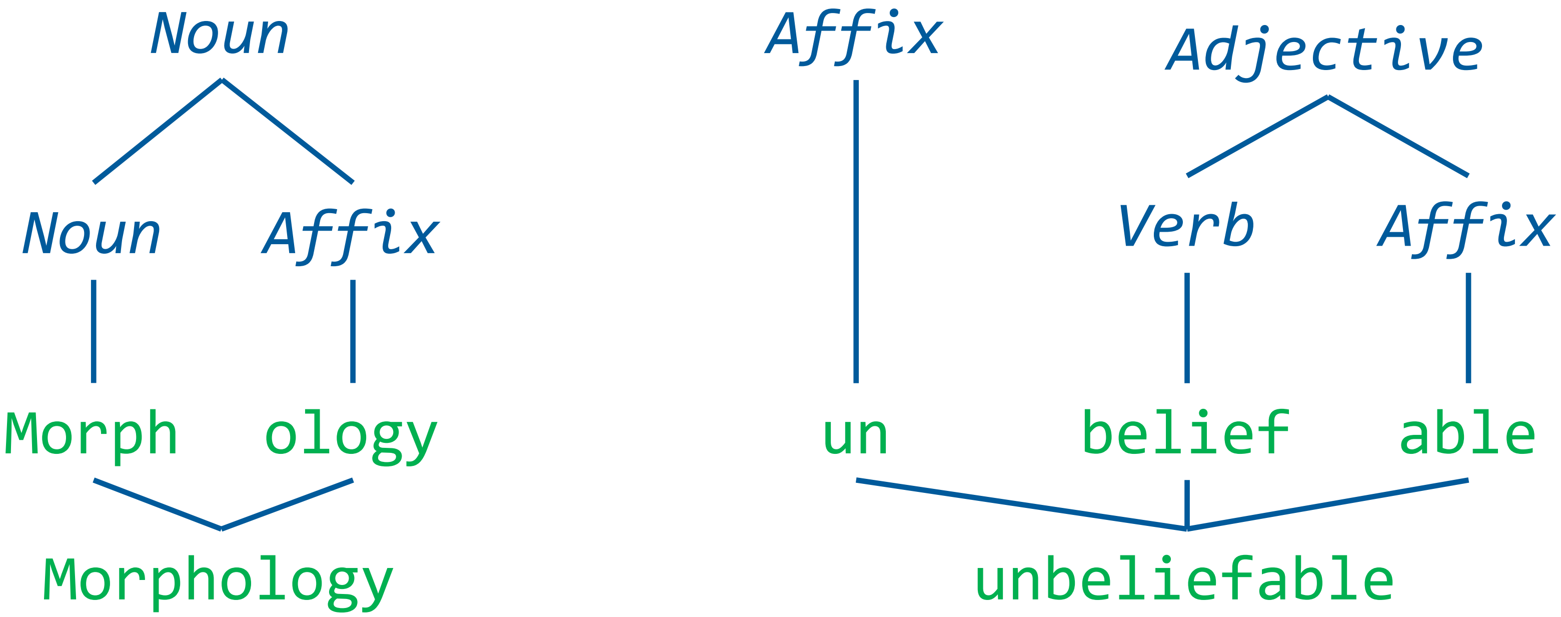


Morphology Tree: Example 1

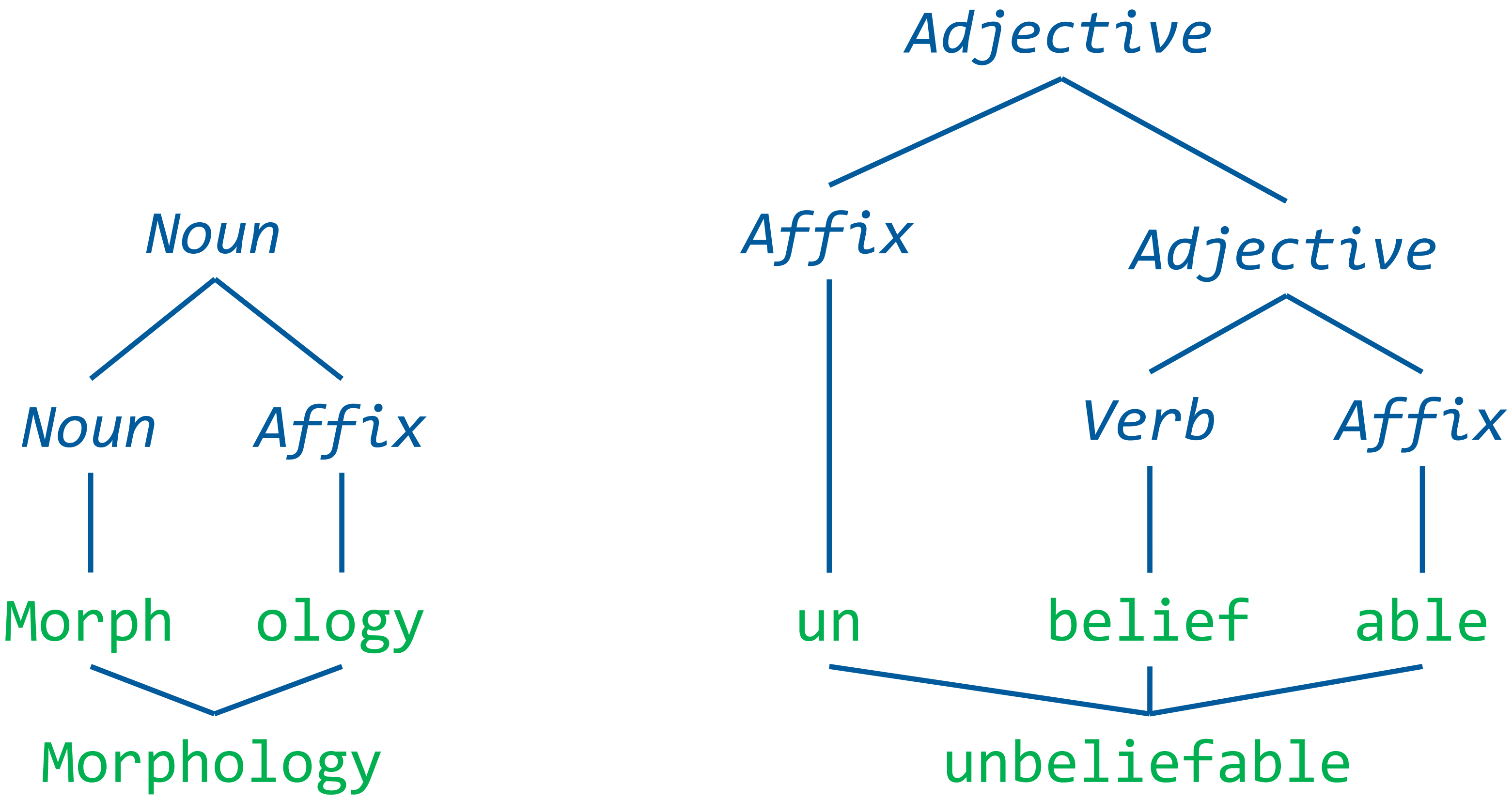




Morphology Tree: Example 1



Morphology Tree: Example 1



Morphology Tree: Example 2

Antidisestablishmentarianism

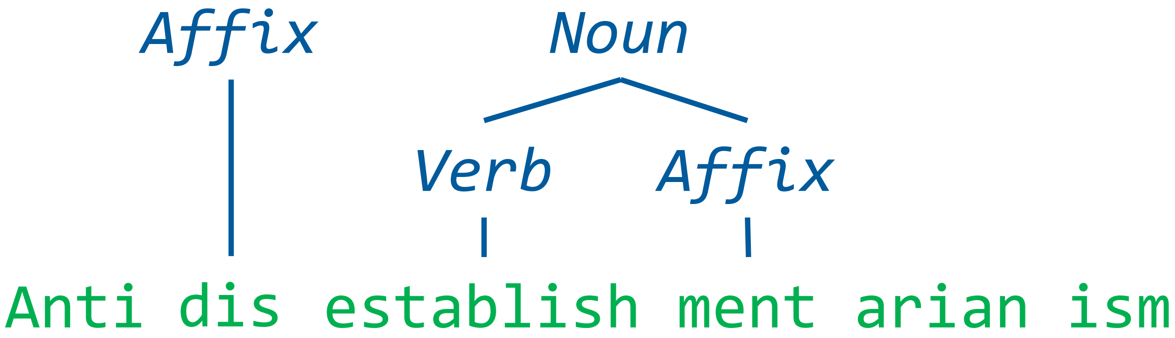
Morphology Tree: Example 2

Anti dis establish ment arian ism

Morphology Tree: Example 2

```
            Verb     Affix
              |         |
Anti dis establish ment arian ism
```

Morphology Tree: Example 2



Affix

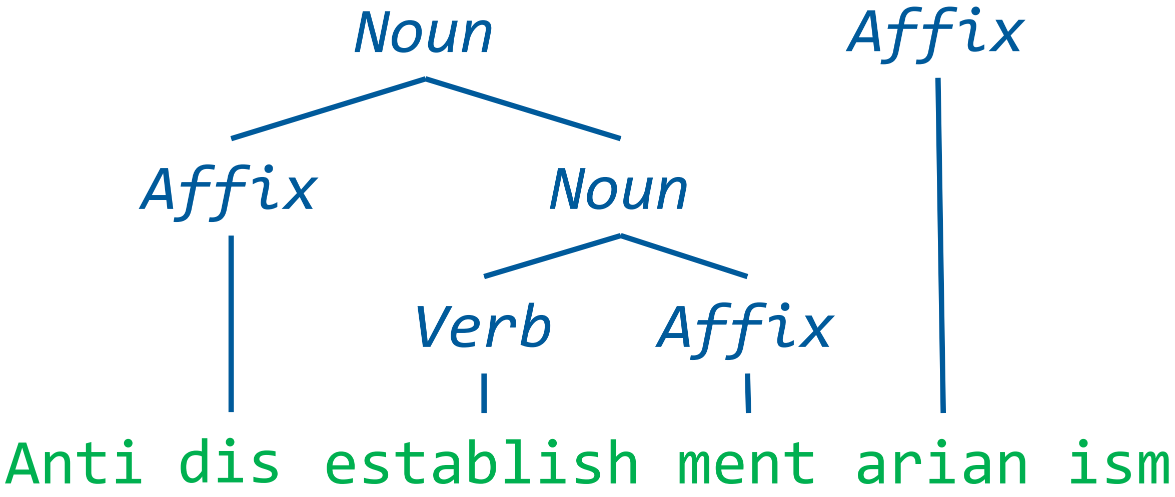
Noun

Verb

Affix

Anti dis establish ment arian ism

Morphology Tree: Example 2



Noun

Affix

Affix

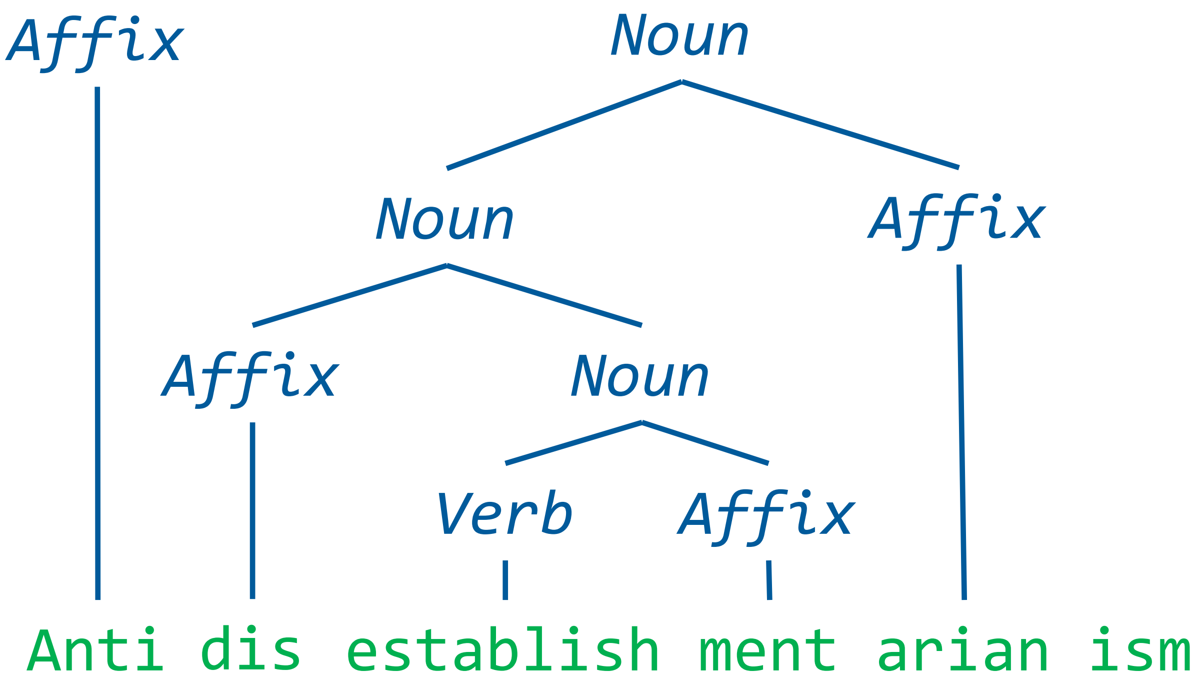
Noun

Verb

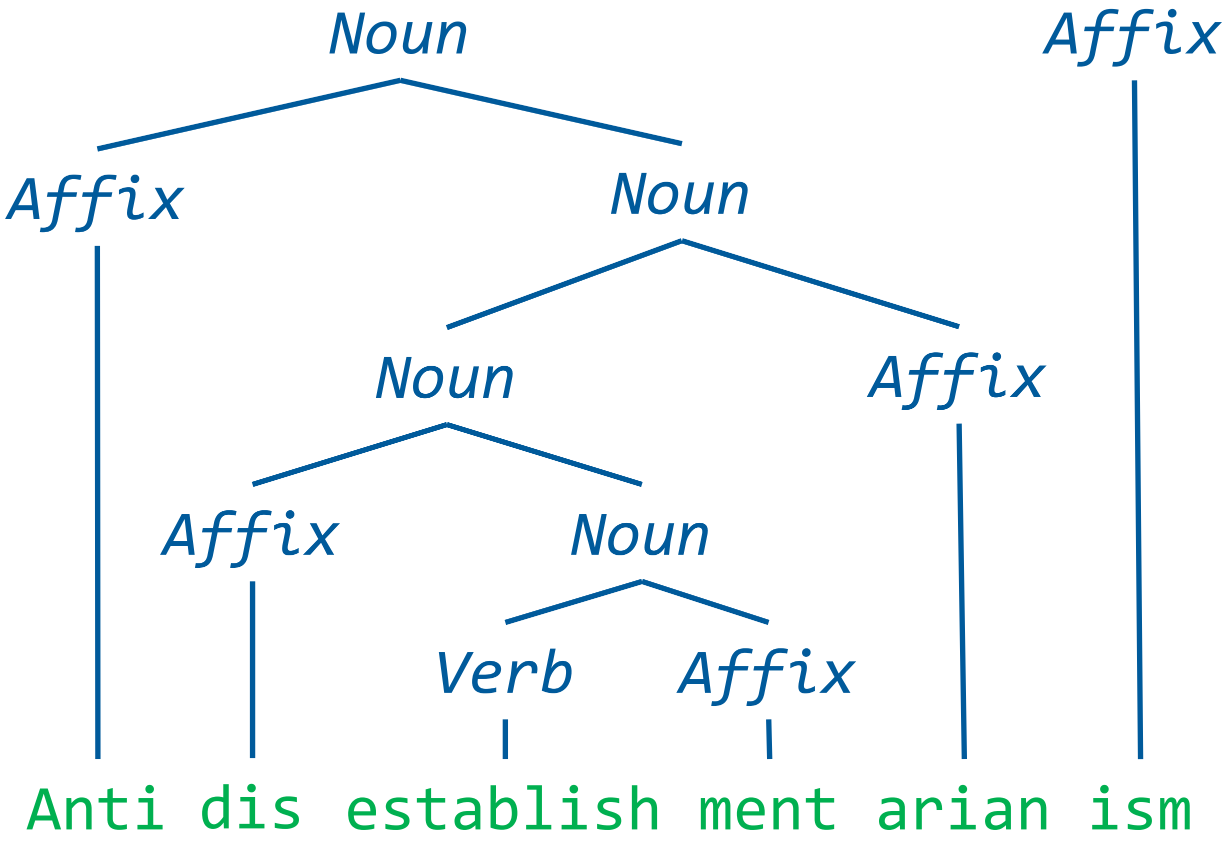
Affix

Anti dis establish ment arian ism

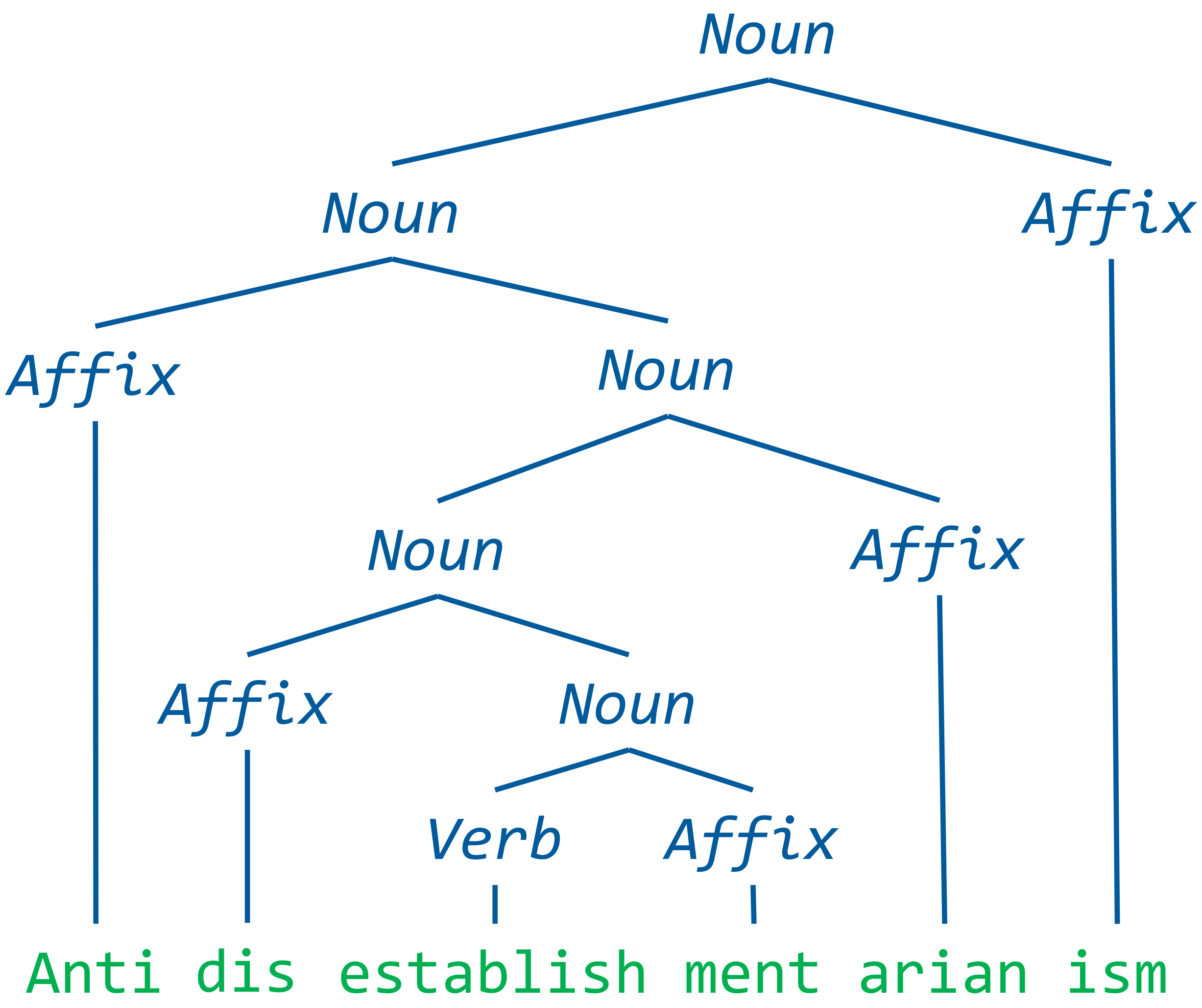
Morphology Tree: Example 2



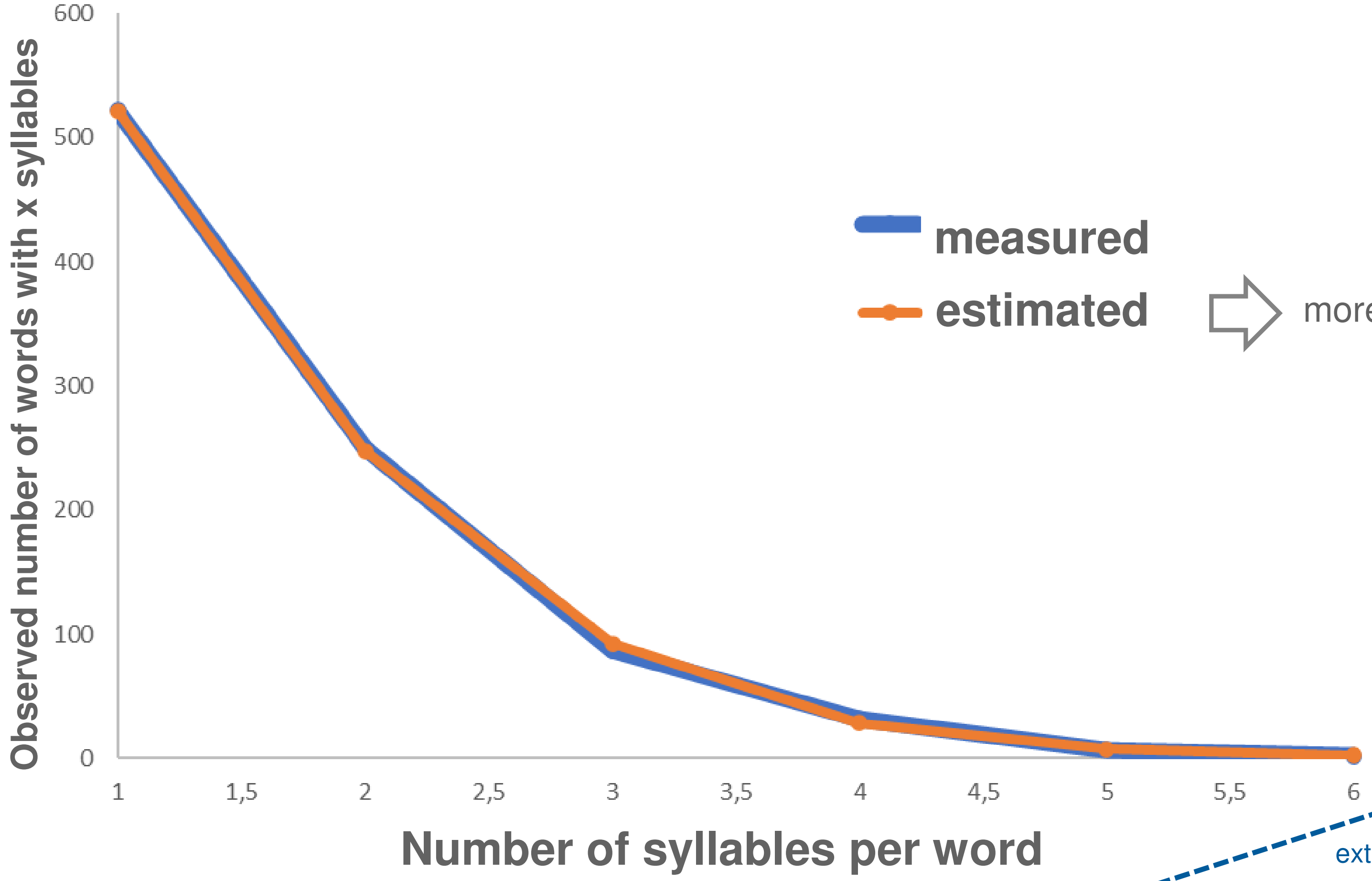
Morphology Tree: Example 2



Morphology Tree: Example 2



Word lengths



more will follow later

extremely rare

Lebensversicherungsgesellschaftsangesteller

See also: Towards a theory of word length distribution

Turkish

Example: “Uygarlastiramadiklarimizdanmissinizcasina”

(behaving) as if you are among those whom we could not civilize

Uygar	las	tir	ama	dik	lar	imiz	dan	mis	siniz	casina
Civilized	become	cause	not able	past	plural	p1pl	abl	past	2pl	as if

Do you know a better example?

Hungarian

Example: “legeslegmegszentségteleníttethetetlenebbjeitekként”

like the most of most undesecratable ones of you or as your most unsanctifiable

https://github.com/oroszgy/awesome-hungarian-nlp#2-datasets

Do you know a better example?

Mandarin Chinese

- Example cases of inflections:

  我(I) ->我们(we)

  他(he) ->他们 (them, plural)

  哥(friend) ->哥们(friends)

- **Adverbial adjective:**

  小心**地**做事 (do things carefully)

- **Adjective form of nouns:**

  可能 (can)

  可能**性** (the possitility)

- **Adverbalized noun :**

  历史 (history)

  历史**上** (in the history)

Lemmatization

Task of determining that two words have the same root, despite their surface differences

What is the basic form of the word?

Before Lemmatization	After Lemmatization
goose	goose
geese	goose
connects	connect
trouble	trouble
troubling	trouble
troubled	trouble
troubles	trouble

am, are, is, be, were, was => be
car, cars, car’s, cars’ => car

⇨ Complex rule-based systems

Stemming

Simpler version of lemmatization in which we mainly just strip suffixes from the end of the word

Porter’s algorithm

- Martin Porter, 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130−137.

„ trace related words to one and the same string“

- Rule-based: https://tartarus.org/martin/PorterStemmer/def.txt
- Tony Kent Strix award in 2000

Example

Input	Output
connect	connect
connected	connect
connections	connect
connects	connect
trouble	troubl
troubled	troubl
troubles	troubl
troublesome	troublesom

Stemming is crude chopping of affixes. It is language dependent
Example: automate(s), automatic – it is reduced to automat.

Porter’s algorithm

Example

forexample compressed and compression are both accepted as equivalent to compress

→

for ***exampl*** ***compress*** and ***compress*** ***ar*** both ***accept*** as ***equival*** to ***compress***

12 words

10 words

Possible Errors

Over-stemming or „false positive“

universal -> ***univers***

university -> ***univers***

universe -> ***univers***

to „univers“

<u>etymologically</u> related but modern meanings are in widely different domains

These are not synonyms, search engine will likely reduce the relevance of the search results.

Under-stemming or „false negative“

alumnus -> ***alumnu***

alumni -> ***alumni***

alumna/***alumna***e -> ***alumna***

This English word keeps Latin morphology, and so these near-synonyms are not conflated.

Stemming algorithms
To minimize both errors

Porter’s algorithm

Determining vocal-consonant-sequences

C := sequence of consonants
V := sequence of vocals
$(.)^m$:= m repetitions of ”.” with $m \geq 0$

$$[C](VC)^m[V]$$

```
        tr ee
        CC VV

          t o
          C V

       w  eb
      C (VC)^1

        an  t
      (VC)^1 C

    tr oubl e
    CC VVCC V
    C (VC)^1 V

    b etw een
    C VCC VVC
    C   (VC)^2

   tr oubl es
   CC VVCC VC
    C   (VC)^2

   pr iv at e
   CC VC VC V
    C (VC)^2 V

w ik ip ed ia
C VC VC VC VV
C   (VC)^3   V
```

https://iq.opengenus.org/porter-stemmer/

Porter’s algorithm

Shortening rules

(condition) S1 -> S2 if **<stem>S1** and **<stem>** satisfies **(condition)** then **<stem>S2**

1 of > 50 rules:

(m > 1) EMENT -> ‘’

```
<stem>S1  = REPLACEMENT
<stem>    = REPLAC
S1        = EMENT

m of <stem>:
    REPLAC
    C VCC VC
    C  (VC)²
    ⇒ m=2 > 1
Shorten with (m > 1) EMENT ->‘‘
    ⇒ REPLACEMENT wird REPLAC
```

Porter’s algorithm

Shortening rules

(condition) S1 -> S2 if **<stem>S1** and **<stem>** satisfies **(condition)** then **<stem>S2**

Example conditions:

*S - the stem ends with S (and similarly for the other letters).
v - the stem contains a vowel.
m=2 TROUBLES, PRIVATE, OATEN, ORRERY.
*d - the stem ends with a double consonant (e.g. -TT, -SS).
*o - the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

Stemming vs. Lemmatization

- **Stemming always shortens the word!**
- **When we apply lemmatization, the word stem does not even need to be the same: (to be, is, was, were)**

Stemming is used most often.

What is a sentence?

A sentence is a self-contained **linguistic** unit consisting of one or more **words.**

The Sentence

"In Germany we use capital letters to mark the beginning of a sentence"

The sentence ends with a punctuation mark.

- Full stop [.]
- Exclamation mark [!]
- Question mark [?]
- Ellipsis [...]

How to mark sentence structure inside compound sentences?

- Comma [,]
- Semicolon [;]
- Dash [;]

What is a sentence?

"There are many different approaches to defining the term "sentence". There are nearly 200 <u>definitions</u> for the term *sentence*"

Kessel, Reimann: *Basiswissen Deutsche Gegenwartssprache*, Fink, Tübingen 2005, ISBN 3-8252-2704-9, S. 1.

Definition: What is a sentence?

"There are many different approaches to defining the term "sentence". There are nearly 200 definitions for the term *sentence*"

Further definitions:

- The sentence as subject and predicate unit
- The sentence as a speech or text element
- The sentence as communicative unit
- ...

Kessel, Reimann: *Basiswissen Deutsche Gegenwartssprache*, Fink, Tübingen 2005, ISBN 3-8252-2704-9, S. 1.

According to „Duden“ (= German spelling dictionary)

> "The sentence is a closed linguistic unit composed of smaller units (words and word groups)."

- **Since there are also sentences with one word (example: "Go!"), such a definition cannot distinguish the sentence from the word.**

According to „Duden“ (= German spelling dictionary)

"The sentence is a closed linguistic unit composed of smaller units (words and word groups)."

- **Since there are also sentences with one word (example: "Go!"), such a definition cannot distinguish the sentence from the word.**

- **It is also unclear what is meant by the term "linguistic unit". A group of words (syntagma) is also a self-contained linguistic unit**

According to „Duden“ (= German spelling dictionary)

"A sentence is a self-contained unit that has been formed according to the rules of syntax."

According to „Duden“ (= German spelling dictionary)

"A sentence is a self-contained unit that has been formed according to the rules of syntax."

This definition is possibly circular, as "syntax" is sometimes regarded as the technical term for sentence theory.

According to „Duden“ (= German spelling dictionary)

"A sentence is a self-contained unit that has been formed according to the rules of syntax."

This definition is possibly circular, as "syntax" is sometimes regarded as the technical term for sentence theory.

- This means, a sentence is a linguistic unit that is a regular sentence according to the doctrine of sentences. To do this, however, one must know what a sentence is. On the other hand, syntactically incorrect structures can also be called sentences.

According to „Duden“ (= German spelling dictionary)

"A sentence is a self-contained unit that has been formed according to the rules of syntax."

This definition is possibly circular, as "syntax" is sometimes regarded as the technical term for sentence theory.

- This means, a sentence is a linguistic unit that is a regular sentence according to the doctrine of sentences. To do this, however, one must know what a sentence is. On the other hand, syntactically incorrect structures can also be called sentences.

Image from cheezburger.com

According to „Duden“ (= German spelling dictionary)

"A sentence is a self-contained unit that has been formed according to the rules of syntax."

This definition is possibly circular, as "syntax" is sometimes regarded as the technical term for sentence theory.

- This means, a sentence is a linguistic unit that is a regular sentence according to the doctrine of sentences. To do this, however, one must know what a sentence is. On the other hand, syntactically incorrect structures can also be called sentences.
- There are also sentences that are not properly formed and are accepted (acceptability despite a lack of (scholastic) grammaticality). Thus, in the case of deliberate violations of selection restrictions:

Example: “Wir sind Papst!” (German: "We are Pope!“)

Image from cheezburger.com

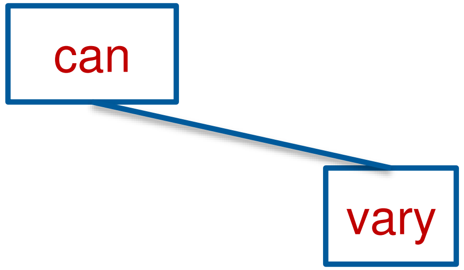
Image from Wikipedia

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."

According to „Duden“ (= German spelling dictionary)

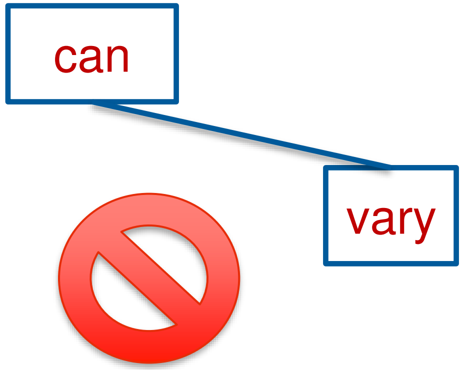
A sentence is a unit consisting of a finite verb and all the clauses required by the verb."



can vary

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."



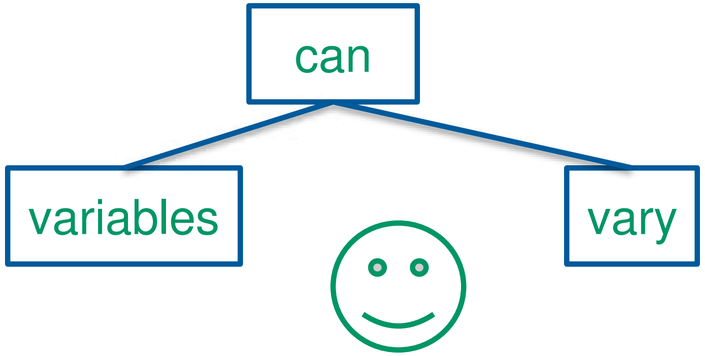
can

vary

can vary

According to „Duden“ (= German spelling dictionary)

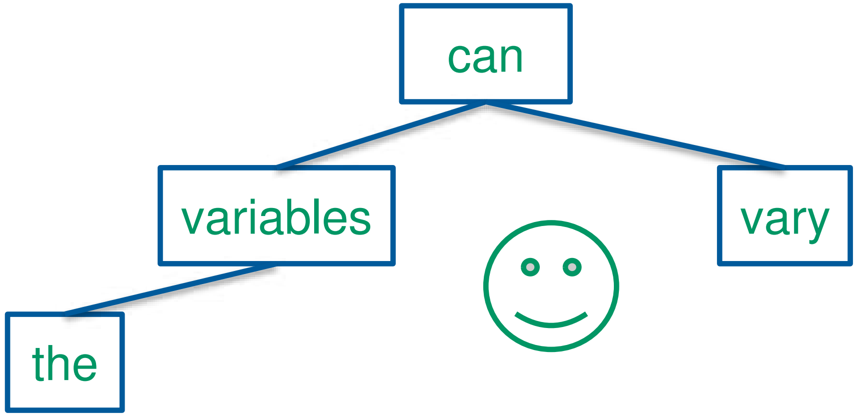
A sentence is a unit consisting of a finite verb and all the clauses required by the verb."



variables
can vary

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."



the variables
can vary

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."

However, *elliptical usage* is also possible, e.g.

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."

However, *elliptical usage* is also possible, e.g.

- She can help with the housework. Nancy can (help with the housework), too.

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."

However, *elliptical usage* is also possible, e.g.

- She can help with the housework. Nancy can (help with the housework), too.
- John can speak seven languages. But Ron can speak only two (languages.)

According to „Duden“ (= German spelling dictionary)

A sentence is a unit consisting of a finite verb and all the clauses required by the verb."

However, *elliptical usage* is also possible, e.g.

- She can help with the housework. Nancy can (help with the housework), too.
- John can speak seven languages. But Ron can speak only two (languages.)
- Lacy can do something about the problem. But I don’t know what(she can do).

See also:
https://en.wikipedia.org/wiki/Ellipsis_(linguistics)

Further Sentence Defintion

“Independent linguistic form that is not contained in a larger linguistic form by a grammatical construction“

Definition according to Bloomfield whose works count as foundation of *American structuralism*.

Further Sentence Defintion

“Independent linguistic form that is not contained in a larger linguistic form by a grammatical construction“

Definition according to Bloomfield whose works count as foundation of *American structuralism*.

- This definition is also possibly circular

Further Sentence Defintion

> “Independent linguistic form that is not contained in a larger linguistic form by a grammatical construction“

Definition according to Bloomfield whose works count as foundation of *American structuralism*.

- This definition is also possibly circular
- Moreover, according to this definition, subordinate clauses are not

sentences, but only clauses.

Further reading

- About *clauses*: https://liberalarts.oregonstate.edu/wlf/what-clause-oregon-state-guide-grammar
- „Sentence and Word“, L. Bloomfield, 1914, https://doi.org/10.2307/282688

Classification of sentences

- **By speech act**
    - declarative sentence: „You are my friend.“
    - Interrogative sentence: „Are you my friend?“
    - prompt sentence: „Be my friend!“

Classification of sentences

- **By speech act**
  - declarative sentence: „You are my friend.“
  - Interrogative sentence: „Are you my friend?“
  - prompt sentence: „Be my friend!“
- **after the verb position of the finite verb in:**
  - begin of sentence (German: „Stirnsatz“)
  - after first clause (German: „Kernsatz“)
  - end of sentence (German: „Spannsatz“)

“[…] those [verbs] inflected for number and person” (c.f. Wikipedia)

Classification of sentences

- **By speech act**
  - declarative sentence: „You are my friend.“
  - Interrogative sentence: „Are you my friend?“
  - prompt sentence: „Be my friend!“

“[…] those [verbs] inflected for number and person” (c.f. Wikipedia)

- **after the verb position of the finite verb in:**
  - begin of sentence (German: „Stirnsatz“)
  - after first clause (German: „Kernsatz“)
  - end of sentence (German: „Spannsatz“)
- **After number and relation of finite verbs in:**
  - Simple sentence: single independent clause, no dependent clause
  - Compound sentence: multiple independent clauses without dependent clauses

Classification of sentences

- **By speech act**
  - declarative sentence: „You are my friend.“
  - Interrogative sentence: „Are you my friend?“
  - prompt sentence: „Be my friend!“
- **after the verb position of the finite verb in:**

  “[…] those [verbs] inflected for number and person” (c.f. Wikipedia)

  - begin of sentence (German: „Stirnsatz“)
  - after first clause (German: „Kernsatz“)
  - end of sentence (German: „Spannsatz“)
- **After number and relation of finite verbs in:**
  - Simple sentence: single independent clause, no dependent clause
  - Compound sentence: multiple independent clauses without dependent clauses
- **Main clause vs. subordinate clause**
- **Syntactic (in)completeness: Anacoluth, ellipsis, fragment, nominal clause, …**

Classification of sentences

- **By speech act**
  - declarative sentence: „You are my friend.“
  - Interrogative sentence: „Are you my friend?“
  - prompt sentence: „Be my friend!“
- **after the verb position of the finite verb in:**
  - begin of sentence (German: „Stirnsatz“)
  - after first clause (German: „Kernsatz“)
  - end of sentence (German: „Spannsatz“)
- **After number and relation of finite verbs in:**
  - Simple sentence: single independent clause, no dependent clause
  - Compound sentence: multiple independent clauses without dependent clauses
- **Main clause vs. subordinate clause**
- **Syntactic (in)completeness: Anacoluth, ellipsis, fragment, nominal clause, …**

“[…] those [verbs] inflected for number and person” (c.f. Wikipedia)

“Cleopatra's nose, had it been shorter, the whole face of the world would have been changed.” (Blaise Pascal)

Spoken Language

- **When speaking, as a rule, at least in German/English/etc., a *short pause* separates a sentence from a preceding one.**

Spoken Language

- **When speaking, as a rule, at least in German/English/etc., a *short pause* separates a sentence from a preceding one.**
- **Sentence melody sometimes depends on the type of sentence (statement, question, request)**

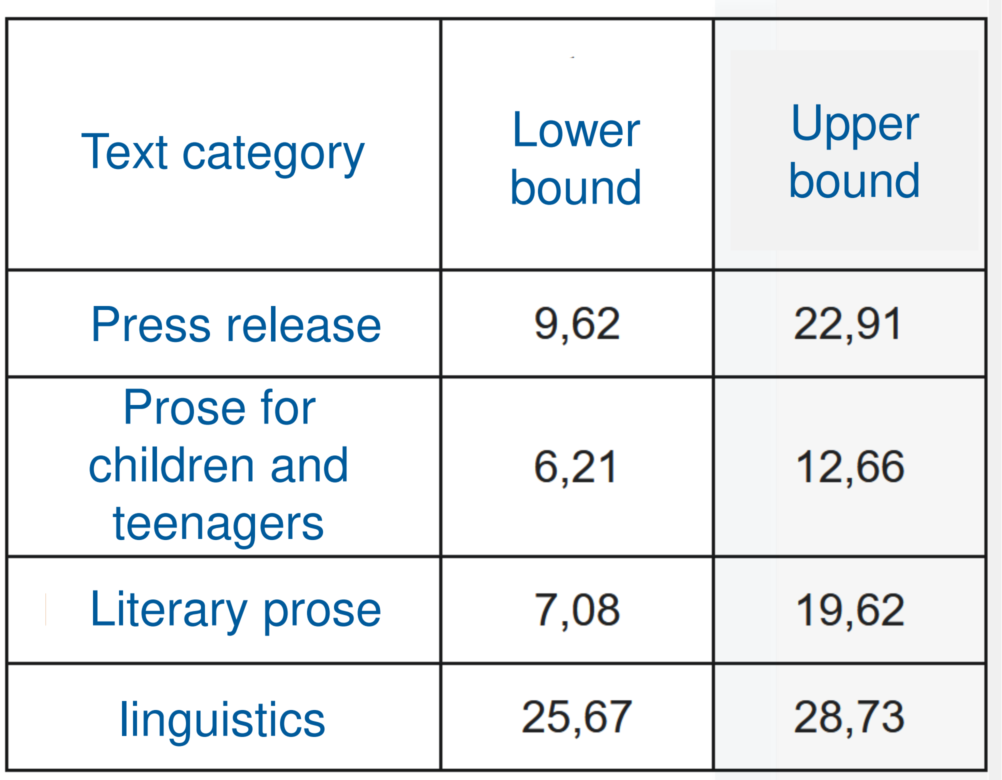
Spoken Language

- **When speaking, as a rule, at least in German/English/etc., a *short pause* separates a sentence from a preceding one.**
- **Sentence melody sometimes depends on the type of sentence (statement, question, request)**
- **A sentence can (usually) be recognized as a unit**

Spoken Language

- **When speaking, as a rule, at least in German/English/etc., a *short pause* separates a sentence from a preceding one.**
- **Sentence melody sometimes depends on the type of sentence (statement, question, request)**
- **A sentence can (usually) be recognized as a unit**
- **The assignment of sentences and their meaning is not always clear**

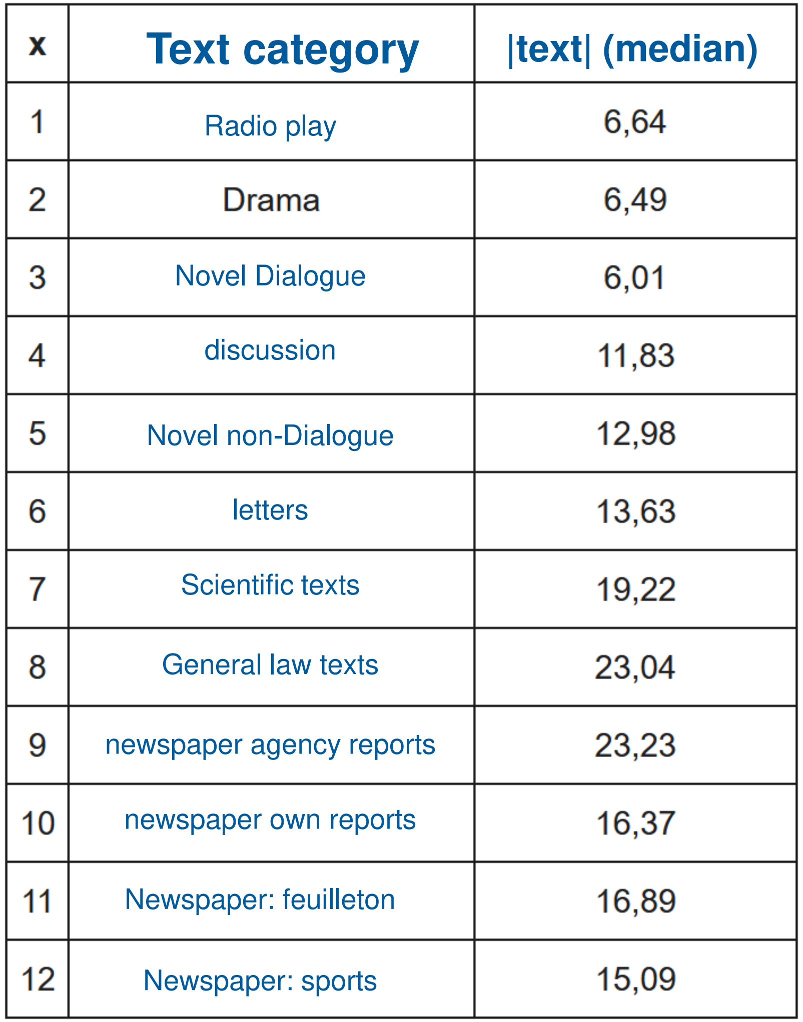
Sentence length in German Literature



Text category	Lower bound	Upper bound
Press release	9,62	22,91
Prose for children and teenagers	6,21	12,66
Literary prose	7,08	19,62
linguistics	25,67	28,73

Karl-Heinz Best: *Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel*. In: *Göttinger Beiträge zur Sprachwissenschaft* 7, 2002, S. 7–31; only the observed values of the record lengths are always given here. All data compiled in the table are based on texts from the 20th century.
https://de.wikipedia.org/wiki/Satzl%C3%A4nge

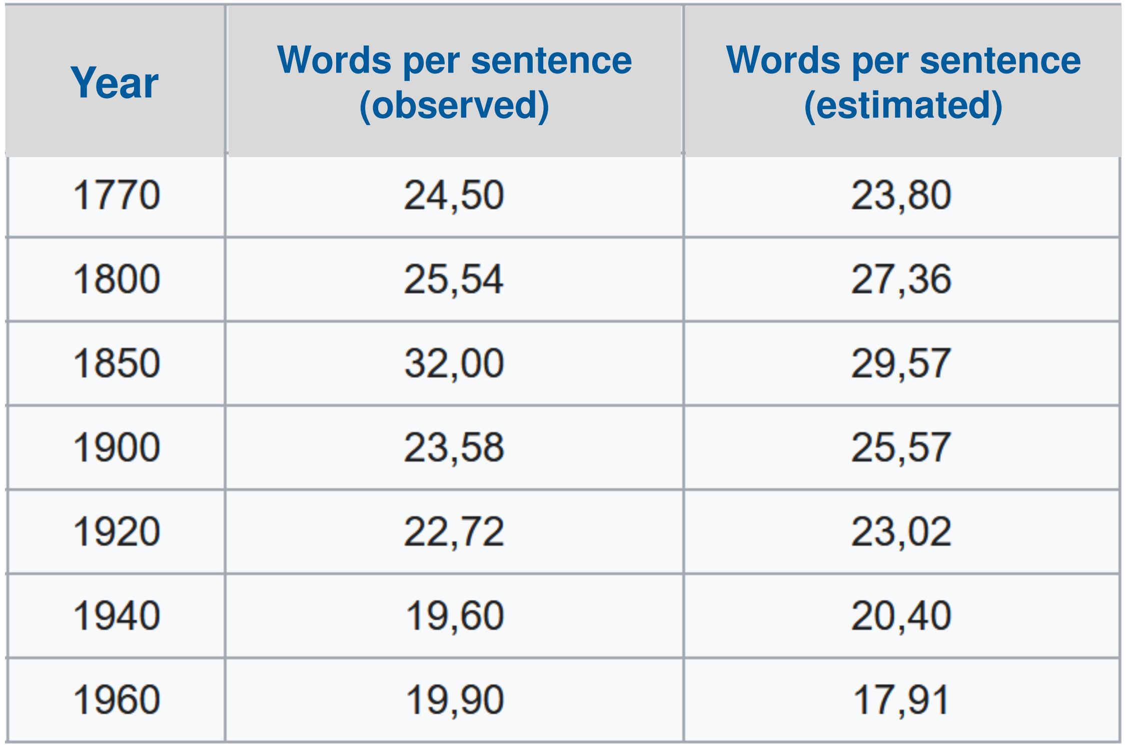
Sentence length in German Literature



x	Text category	|text| (median)
1	Radio play	6,64
2	Drama	6,49
3	Novel Dialogue	6,01
4	discussion	11,83
5	Novel non-Dialogue	12,98
6	letters	13,63
7	Scientific texts	19,22
8	General law texts	23,04
9	newspaper agency reports	23,23
10	newspaper own reports	16,37
11	Newspaper: feuilleton	16,89
12	Newspaper: sports	15,09

Ursula Pieper: *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse.* Narr, Tübingen 1979, Seite 50. ISBN 3-87808-355-6.

Piotrowski‘s Law



Year	Words per sentence (observed)	Words per sentence (estimated)
1770	24,50	23,80
1800	25,54	27,36
1850	32,00	29,57
1900	23,58	25,57
1920	22,72	23,02
1940	19,60	20,40
1960	19,90	17,91

Karl-Heinz Best: *Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel*. In: *Göttinger Beiträge zur Sprachwissenschaft* 7, 2002, Seite 7–31, zur Entwicklung der Satzlängen Seite 22–27, table on page 25, corrected.
https://de.wikipedia.org/wiki/Piotrowski-Gesetz