What is a vocabulary?

List of all known words of a language

Vocabulary

- **List of all words**

Vocabulary

- **List of all words**
- **When pronunciation is added, one also speaks of "lexicons".**

Vocabulary

Speech recognition
Speech synthesis

- **List of all words**
- **When pronunciation is added, one also speaks of "lexicons".**

Vocabulary

Speech recognition
Speech synthesis

- **List of all words**
- **When pronunciation is added, one also speaks of "lexicons".**
- **Most often divided into classes:**
  - to be, was, were
  - car, cars

Vocabulary

Speech recognition
Speech synthesis

- **List of all words**
- **When pronunciation is added, one also speaks of "lexicons".**
- **Most often divided into classes:**
  - to be, was, were
  - car, cars
- **with semantic paraphrasing (wordnet):**

Vocabulary

Speech recognition
Speech synthesis

- **List of all words**
- **When pronunciation is added, one also speaks of "lexicons".**
- **Most often divided into classes:**
  - to be, was, were
  - car, cars
- **with semantic paraphrasing (see wordnet):**
  - car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) "he needs a car to get to work"
  - car, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of railroad) "three cars had jumped the rails"
  - car, gondola (the compartment that is suspended from an airship and that carries personnel and the cargo and the power plant)
  - car, elevator car (where passengers ride up and down) "the car was on the top floor"
  - cable car, car (a conveyance for passengers or freight on a cable railway) "they took a cable car to the top of the mountain"

Information Retrieval

Text Preprocessing

1. **Tokenized (Word segmentation)**
2. **Normalization: (comparability)**
   - Normalisieren
   - Groß-/Kleinschreibung
   - Morphology
   - Lemmatisierung/Stemming
3. **Sentence Segmentation**

What is a vocuabulary?
Frequency of words?

Zipf‘s Law

Rank vs. Frequency

What is it about?

- **George Kingsley Zipf (1902-1950)**
- **Relation between the frequency-rank of a word and its frequency**

What is it about?

- **George Kingsley Zipf (1902-1950)**
- **Relation between the frequency-rank of a word and its frequency**
- n **:= Position within the ranking, e. g. the word „the“ is on rank 1 in English**

What is it about?

- George Kingsley Zipf (1902-1950)
- Relation between the frequency-rank of a word and its frequency
- n := Position within the ranking, e. g. the word „the“ is on rank 1 in English
- $f(n)$:= frequency of word with rank n

What is it about?

- **George Kingsley Zipf (1902-1950)**
- **Relation between the frequency-rank of a word and its frequency**
- n **:= Position within the ranking, e. g. the word „the“ is on rank 1 in English**
- $f(n)$ **:= frequency of word with rank n**
- s **:= normalization factor close to 1**

$$f(n) \propto \frac{1}{n^s}$$

Ferrer i Cancho, Ramon, and Ricard V Sole. “Least effort and the origins of scaling in human language.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 100,3 (2003): 788-91. doi:10.1073/pnas.0335980100

What is it about?

- George Kingsley Zipf (1902-1950)
- Relation between the frequency-rank of a word and its frequency
- n := Position within the ranking, e. g. the word „the“ is on rank 1 in English
- $f(n)$:= frequency of word with rank n
- s := normalization factor close to 1

$$f(n) \propto \frac{1}{n^s}$$

What does this symbol mean?

Ferrer i Cancho, Ramon, and Ricard V Sole. “Least effort and the origins of scaling in human language.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 100,3 (2003): 788-91. doi:10.1073/pnas.0335980100

What is it about?

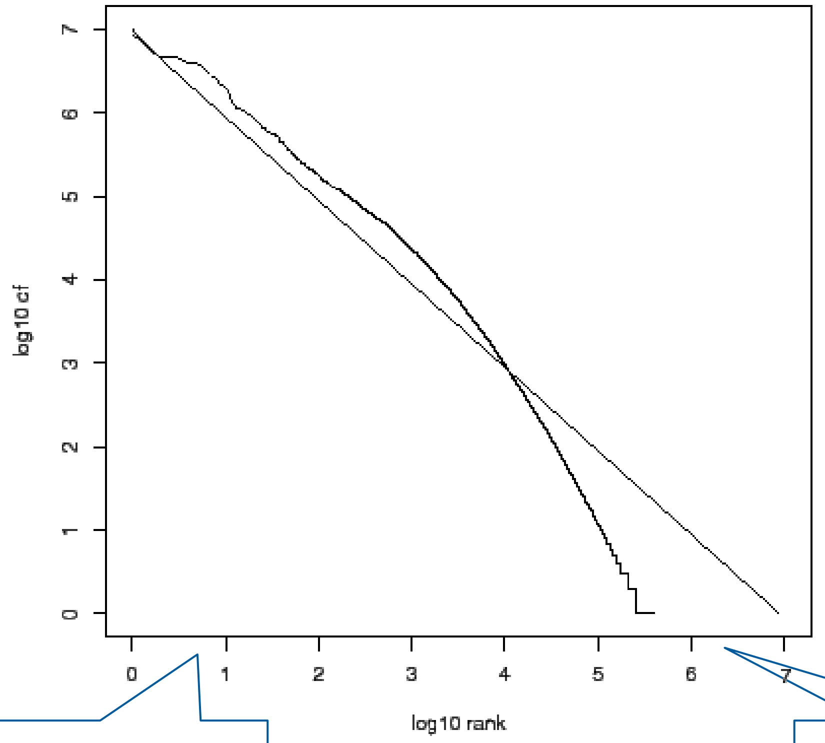
- George Kingsley Zipf (1902-1950)
- Relation between the frequency-rank of a word and its frequency
- n := Position within the ranking, e. g. the word „the“ is on rank 1 in English
- $f(n)$:= frequency of word with rank n
- s := normalization factor close to 1

$$f(n) \propto \frac{1}{n^s} \quad \Leftrightarrow \quad \exists c \in \mathbb{R} \setminus \{0\}\colon\ f(n) = \frac{c}{n^s}$$

Ferrer i Cancho, Ramon, and Ricard V Sole. “Least effort and the origins of scaling in human language.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 100,3 (2003): 788-91. doi:10.1073/pnas.0335980100

Visualization of Zipf‘s Law: Reuters-RCV1 corpus

$$f(n) = \frac{c}{n^s}$$



Most common words

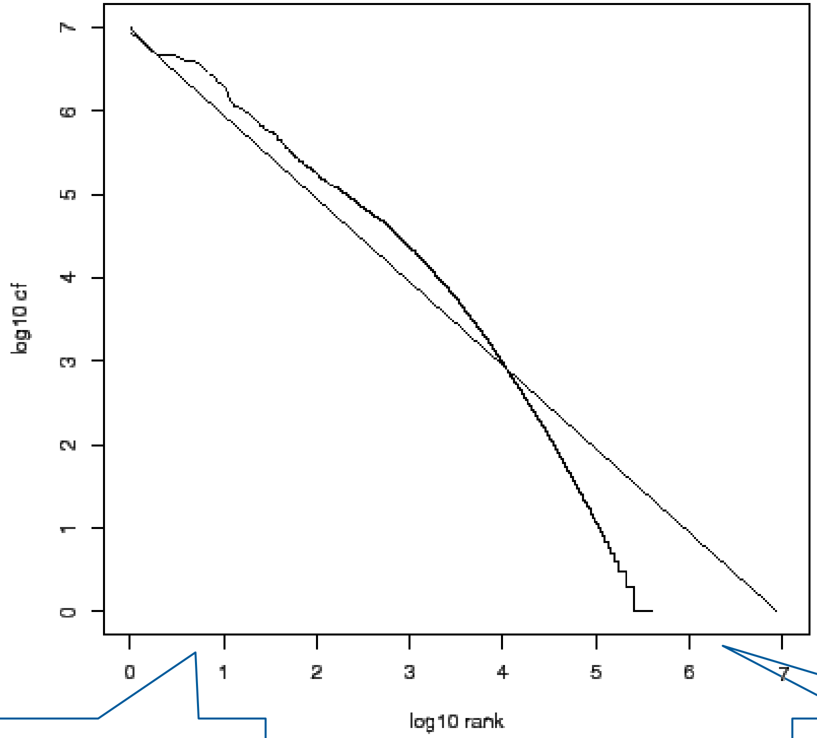
Less common words

Image source: https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html

Visualization of Zipf's Law: Reuters-RCV1 corpus

$$f(n) = \frac{c}{n^s}$$

$$\log(f(n)) = \log c - s \log(n)$$



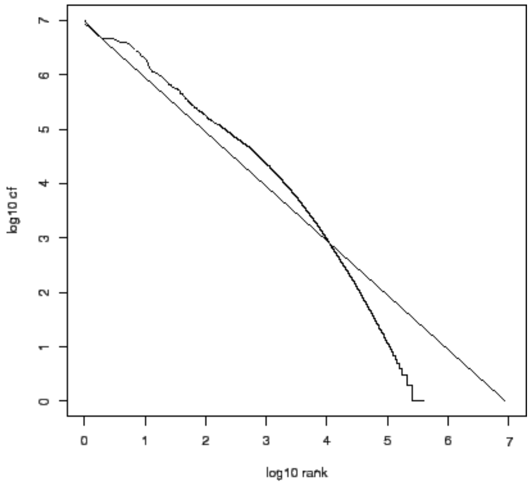
Most common words

Less common words

Image source: https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html

Notes on Zipf‘s Law (1/3)

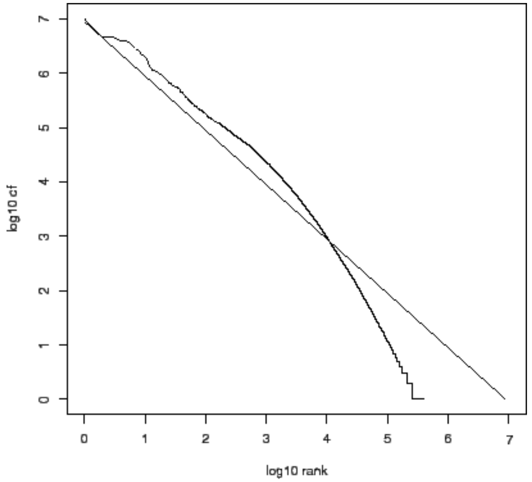
- **Possible Explanation: „Principle of least effort“**



Notes on Zipf‘s Law (1/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

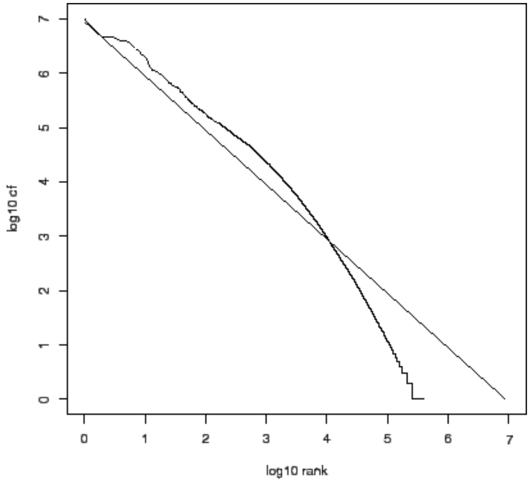
  ⇔ many short words and polysemy



Notes on Zipf‘s Law (1/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

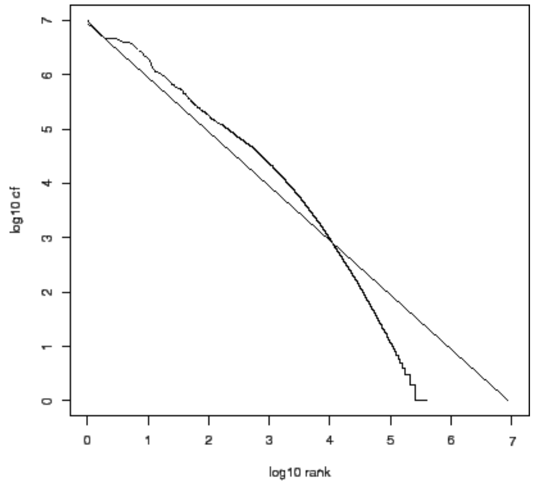
  ⇔ many short words and polysemy
  - Listeners don‘t like ambiguity ⇔ vocab size large



Notes on Zipf‘s Law (2/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

  ⇔ many short words and polysemy
  - Listeners don‘t like ambiguity ⇔ vocab size large

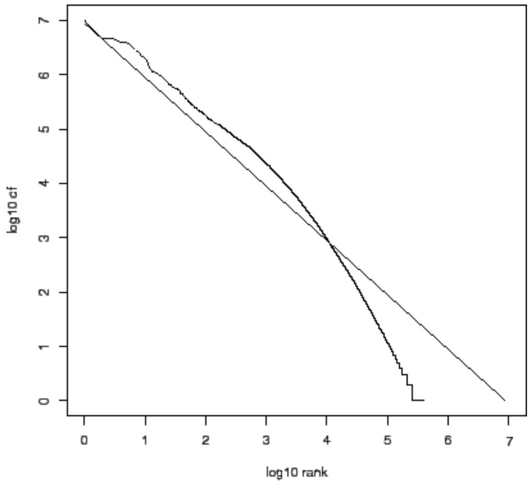


- **Holds for other „rankable“ quantities:**

Notes on Zipf‘s Law (2/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

  ⇔ many short words and polysemy
  - Listeners don‘t like ambiguity ⇔ vocab size large

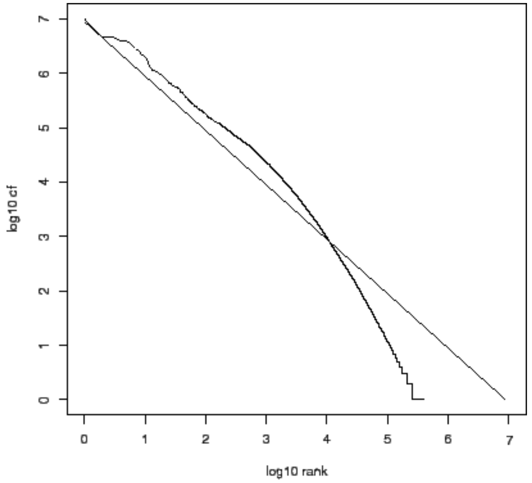


- **Holds for other „rankable“ quantities:**
  - websites ranked by daily visits,

Notes on Zipf‘s Law (2/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

    ⇔ many short words and polysemy
  - Listeners don‘t like ambiguity ⇔ vocab size large

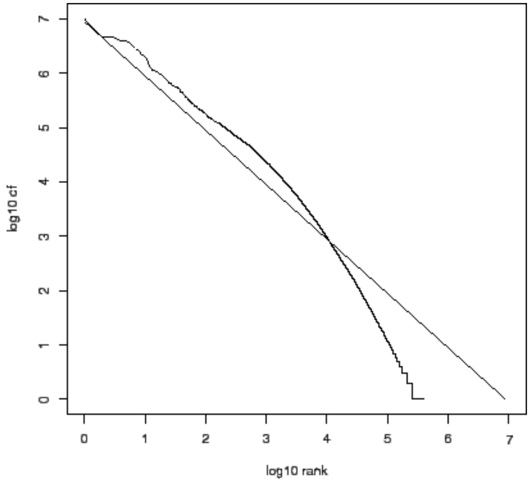


- **Holds for other „rankable“ quantities:**
  - websites ranked by daily visits,
  - cities w.r.t. inhabitants, …

Notes on Zipf‘s Law (2/3)

- **Possible Explanation: „Principle of least effort“**
  - Speakers don‘t like a 1:1 vocabulary and less effort in general

  ⇔ many short words and polysemy
  - Listeners don‘t like ambiguity ⇔ vocab size large



- **Holds for other „rankable“ quantities:**
  - websites ranked by daily visits,
  - cities w.r.t. inhabitants, …

Question:
words <-> meanings => **What about the distribution of meanings?**

References

- **“Human Behavior And The Principle Of Least Effort”, G.K. Zipf, 1949, *https://archive.org/details/in.ernet.dli.2015.90211/mode/2up***
- ***Article „Zipf, Power-laws, and Pareto - a ranking tutorial”*, available online**

Notes on Zipf‘s Law (3/3)

Remember: $f(n) \propto \frac{1}{n^s}$

- **Related to *power laws*, like Zeta distribution, defined for positive integers** $\mathrm{n} \geq 1$ and $\mathrm{s} \in \mathbb{R}$

Notes on Zipf‘s Law (3/3)

Remember: $f(n) \propto \frac{1}{n^s}$

- **Related to *power laws*, like Zeta distribution, defined for positive integers** $\mathrm{n} \geq 1$ and $\mathrm{s} \in \mathbb{R}$
  - **PMF** $P(X = n) = \frac{n^{-s}}{\zeta(s)}$

Notes on Zipf‘s Law (3/3)

Remember: $f(n) \propto \frac{1}{n^s}$

- **Related to *power laws*, like Zeta distribution, defined for positive integers** $\mathrm{n} \geq 1$ and $\mathrm{s} \in \mathbb{R}$
  - **PMF** $P(X = n) = \frac{n^{-s}}{\zeta(s)}$
  - **CDF** $P(X \leq m) = \frac{H_{m,s}}{\zeta(s)}$, with generalized *harmonic number* $H_{m,s} \coloneqq \sum_{n=1}^{m} \frac{1}{n^s}$

    and *Riemann-Zeta function* $\zeta(s) \coloneqq \lim_{m \to \infty} H_{m,s}$

Notes on Zipf‘s Law (3/3)

Remember: $f(n) \propto \frac{1}{n^s}$

- **Related to *power laws*, like Zeta distribution, defined for positive integers** $n \geq 1$ and $s \in \mathbb{R}$
  - **PMF** $P(X = n) = \frac{n^{-s}}{\zeta(s)}$
  - **CDF** $P(X \leq m) = \frac{H_{m,s}}{\zeta(s)}$, with generalized *harmonic number* $H_{m,s} := \sum_{n=1}^{m} \frac{1}{n^s}$

    and *Riemann-Zeta function* $\zeta(s) := \lim_{m \to \infty} H_{m,s}$

Question: For which $s \in \mathbb{R}$ is $\zeta(s) < \infty$?

Note: $\mathbb{P}$ = set of prime numbers: $\zeta(s) = \prod_{p \in \mathbb{P}} \frac{1}{1-p^{-s}}$

Notes on Zipf's Law (3/3)

Remember: $f(n) \propto \frac{1}{n^s}$

- **Related to *power laws*, like Zeta distribution, defined for positive integers** $n \geq 1$ and $s \in \mathbb{R}$
  - **PMF** $P(X = n) = \frac{n^{-s}}{\zeta(s)}$
  - **CDF** $P(X \leq m) = \frac{H_{m,s}}{\zeta(s)}$, with generalized *harmonic number* $H_{m,s} := \sum_{n=1}^{m} \frac{1}{n^s}$

    and *Riemann-Zeta function* $\zeta(s) := \lim_{m \to \infty} H_{m,s}$

Question: For which $s \in \mathbb{R}$ is $\zeta(s) < \infty$?

Note: $\mathbb{P}$ = set of prime numbers: $\zeta(s) = \prod_{p \in \mathbb{P}} \frac{1}{1-p^{-s}}$

Further Reading

- **"Human Behavior And The Principle Of Least Effort", G.K. Zipf, 1949, *https://archive.org/details/in.ernet.dli.2015.90211/mode/2up***
- ***Article „Zipf, Power-laws, and Pareto - a ranking tutorial"*, available online**
- **Reuters-RCV1 corpus and Zipf's Law: https://nlp.stanford.edu/IR-book/html/htmledition/zipfs-law-modeling-the-distribution-of-terms-1.html**

Text Preprocessing

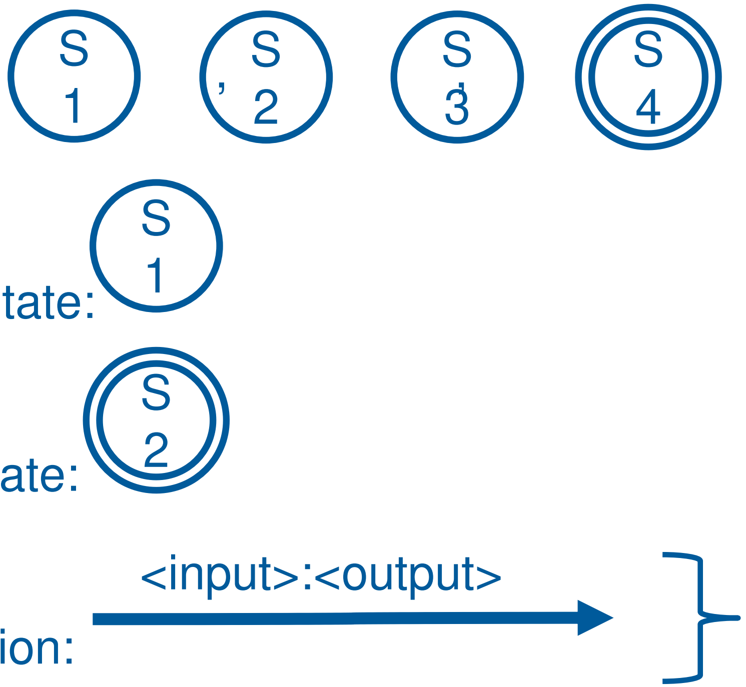
1. **Tokenized (Word segmentation)**
2. **Normalization: (comparability)**
   - Normalisieren
   - Groß-/Kleinschreibung
   - Morphology
   - Lemmatisierung/Stemming
3. **Sentence Segmentation**

What is a vocuabulary?
Frequency of words?
How to implement those steps?

Finite State Transducer

Translate strings

Definition



States:
,

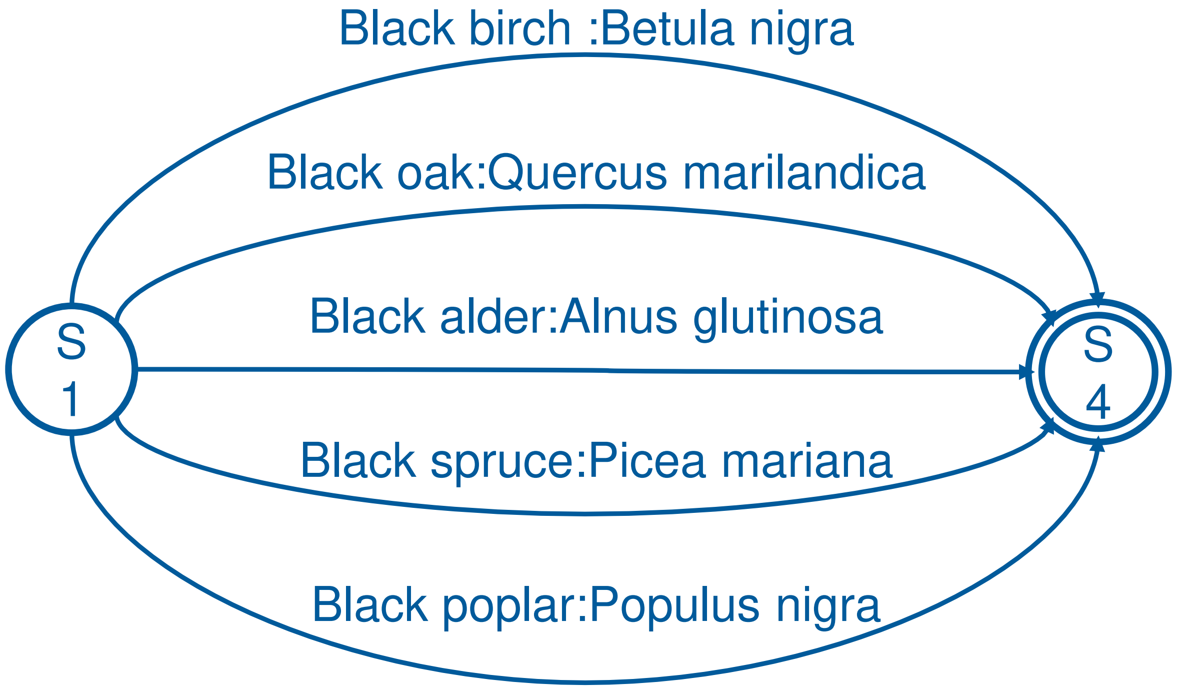
Start-State:

End-State:

Transition:

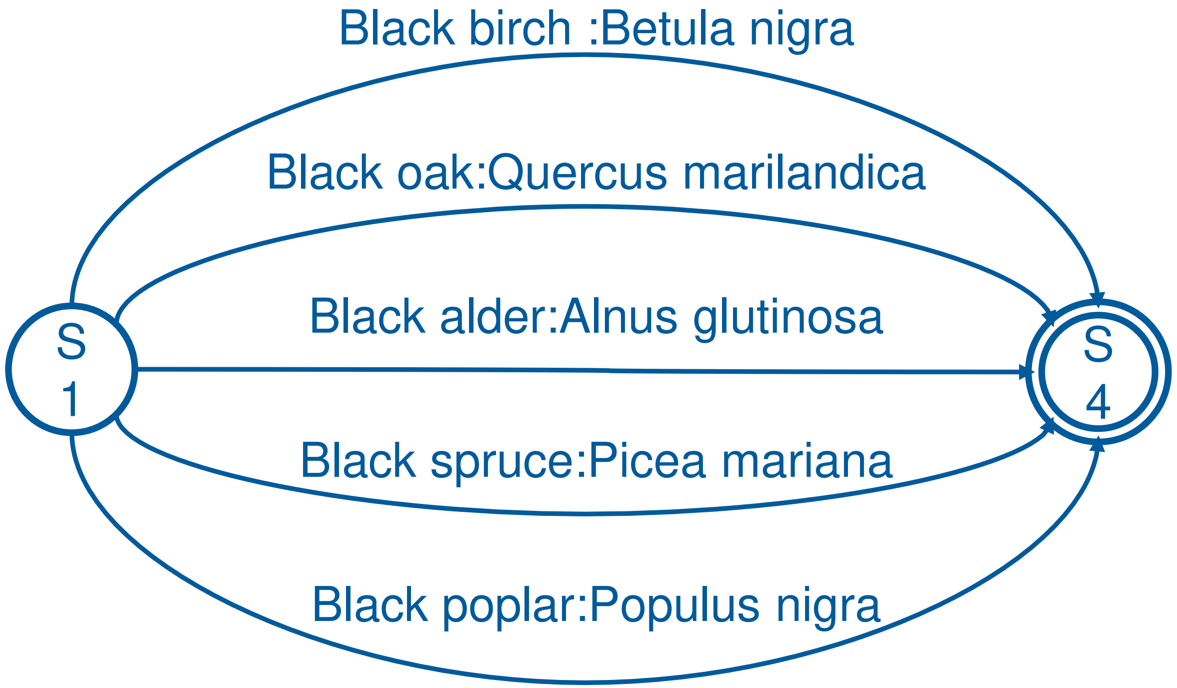
The only difference to the deterministic finite accepter

Example: FST


Black birch :Betula nigra
Black oak:Quercus marilandica
Black alder:Alnus glutinosa
S
1
S
4
Black spruce:Picea mariana
Black poplar:Populus nigra

meaningful?

Example: FST


Black birch :Betula nigra
Black oak:Quercus marilandica
S
1
Black alder:Alnus glutinosa
S
4
Black spruce:Picea mariana
Black poplar:Populus nigra

Doable, but perhaps better broken down into "single characters" per transition...

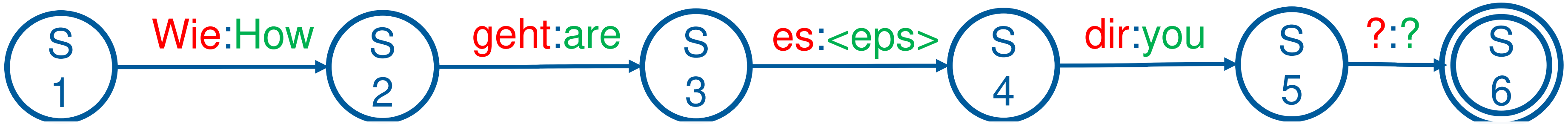
Special Symbol

- **Remember: ε or λ referred to *empty* symbol in last week‘s lecture. Here, it‘s denoted by** „<eps>“
- **<eps> in der Ausgabe:**

- in der Eingabe:

Special Symbol

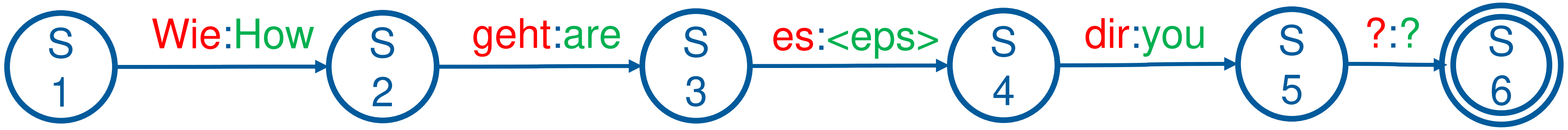
- **Remember: ε or λ referred to *empty* symbol in last week‘s lecture. Here, it‘s denoted by** „<eps>“
- **<eps> in output:**



- **<eps> in input:**

Special Symbol

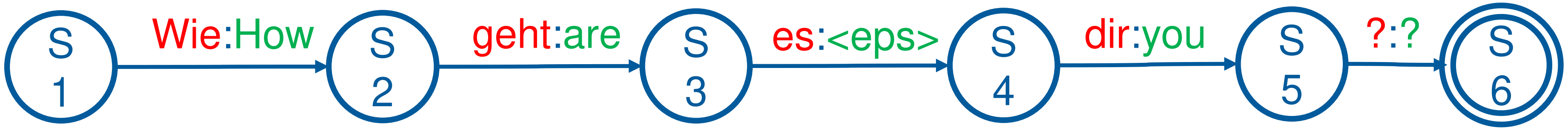
- **Remember: *ε* or *λ* referred to *empty* symbol in last week‘s lecture. Here, it‘s denoted by** „<eps>“
- **<eps> in output:**



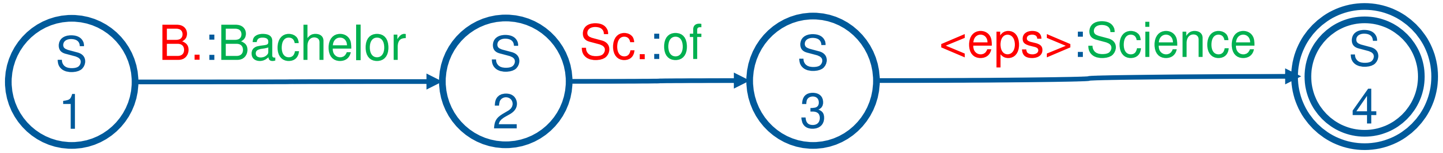
- **<eps> in input:**

Special Symbol

- **Remember: ε or λ referred to *empty* symbol in last week‘s lecture. Here, it‘s denoted by** „<eps>“
- **<eps> in output:**

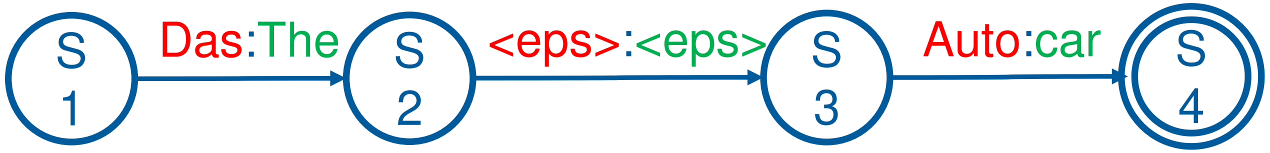


- **<eps> in input:**



Special Symbol

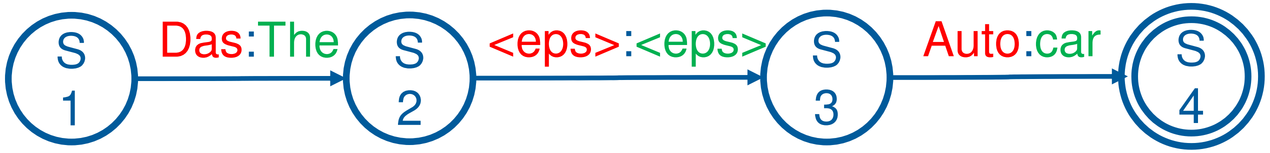
- **<eps> in input and output?**



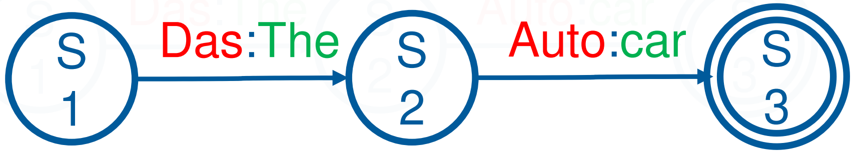
…is equivalent to:

Special Symbol

- **<eps> in input and output?**



…is equivalent to:



Summary

- **Almost the same as a deterministic finite accepter**
- **With an empty symbol (<eps>) an FST quickly becomes non-deterministic**

  (Refresh: every NFA can be transformed into an equivalent DFA)
- **Input and output symbol are considered as one symbol and then the automaton can be minimized as a DFA**

How to implement those steps?

We have a huge data set that we have to preprocess in an automated way.

1. **Tokenization (Word Segmentation)**
2. **Normalization:**
   - Normalizing
   - Upper / lower casing
   - Morphology
   - Lemmatization / Stemming

How to implement those steps?

Food was pretty good, fast, and we met some FBLA members from other states! Wednesday: Well, I was supposed to go River Rafting but if I didn't tell you, I'm not allowed to go (don't ask, LOL). So, the whole AZ Delegation had to get up at 5:30 in the morning and board the buses at 6:00. After they left, I just took a nice long nap (because I went to bed at about 2 AM the last night). By the time it was 11:00 AM they already came back and I woke up, got some lunch and now I am here typing this lovely post. For later today, the opening session starts (which means the actual FBLA Conference starts) and the opening session is about 3 hours long... watch me try to stay awake. I hope I was able to entertain you with this lovely post! And I'll update again in a couple days. Hope you're having a great week! 11,July,2004 Yipee... I finally finished all my packing... at least I think. I still have a few bits and pieces before I'm set. If you forgot what I am packing for, here's a short description: I will be attending the FBLA (Future Business Leaders of America) National Leadership Conference which will be held in Denver, Colorado. It will be held from July 12th to the 18th. I'll remember to take my camera and take some pictures for all of you to see! And if the hotel is really good... I can post some entries while I'm there using their "Business Center". Basically a business center has a copy machine, some computers you can use for internet, etc. for the guests of the hotel. You guys are still free to call my cell at any time. But if I am in the middle of the conference, I won't answer it; so just leave a message and I'll get back to you. I hope all of you have a great week and don't forget about me! 11,July,2004 Welcome to my new blog! Now, I can make posts from anywhere and not just from home. I hope all of you like the new one! 24,July,2004

Scraped blog: 3899990.male.15.Student.Leo.xml

Normalization without stemming

Food was pretty good, fast, and we met some FBLA members from other states! Wednesday: Well, I was supposed to go River Rafting but if I didn't tell you, I'm not allowed to go (don't ask, LOL). So, the whole AZ Delegation had to get up at 5:30 in the morning and board the buses at 6:00. After they left, I just took a nice long nap (because I went to bed at about 2 AM the last night). By the time it was 11:00 AM they already came back and I woke up, got some lunch and now I am here typing this lovely post. For later today, the opening session starts (which means the actual FBLA Conference starts) and the opening session is about 3 hours long... watch me try to stay awake. I hope I was able to entertain you with this lovely post! And I'll update again in a couple days. Hope you're having a great week! 11,July,2004 Yipee... I finally finished all my packing... at least I think. I still have a few bits and pieces before I'm set. If you forgot what I am packing for, here's a short description: I will be attending the FBLA (Future Business Leaders of America) National Leadership Conference which will be held in Denver, Colorado. It will be held from July 12th to the 18th. I'll remember to take my camera and take some pictures for all of you to see! And if the hotel is really good... I can post some entries while I'm there using their "Business Center". Basically a business center has a copy machine, some computers you can use for internet, etc. for the guests of the hotel. You guys are still free to call my cell at any time. But if I am in the middle of the conference, I won't answer it; so just leave a message and I'll get back to you. I hope all of you have a great week and don't forget about me! 11,July,2004 Welcome to my new blog! Now, I can make posts from anywhere and not just from home. I hope all of you like the new one! 24,July,2004

Scraped blog: 3899990.male.15.Student.Leo.xml

Normalization without stemming

Food was pretty good, fast, and we met some FBLA members from other states!

~~Wednesday:~~ Well, I was supposed to go River Rafting but if I didn't tell you, I'm not allowed to go ~~(don't ask, LOL).~~

So, the whole AZ Delegation had to get up at 5:30 in the morning and board the buses at 6:00.

After they left, I just took a nice long nap ~~(because I went to bed at about 2 AM the last night)~~.

By the time it was 11:00 AM they already came back and I woke up, got some lunch and now I am here typing this lovely post.

For later today, the opening session starts ~~(which means the actual FBLA Conference starts)~~ and the opening session is about 3 hours long... watch me try to stay awake.

I hope I was able to entertain you with this lovely post!

And I'll update again in a couple days.

Hope you're having a great week! ~~11,July,2004 Yipee...~~

I finally finished all my packing... ~~at least I think.~~

...

TODO: replace FBLA with Future Business Leaders of America

Scraped blog: 3899990.male.15.Student.Leo.xml

Ok, but for text preprocessing?

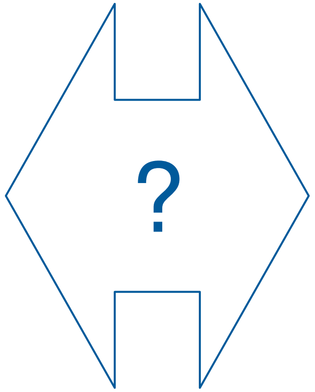
Input: Text as it was written

Output after application of an FST:

- **tokenized**
- **normalized, punctuation removed, ..**
- **Words in their stem form**

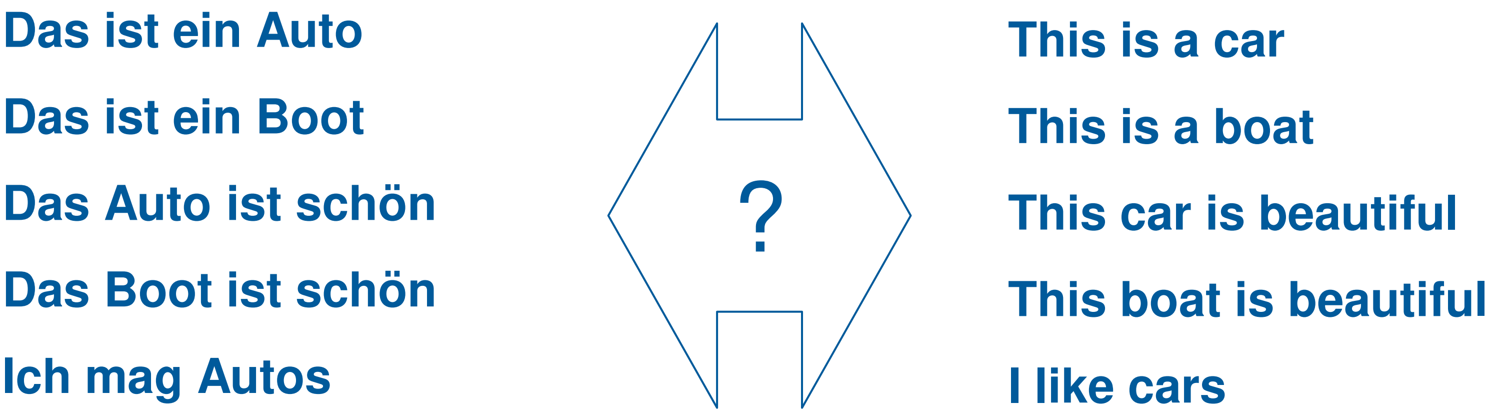
Putting quantities in relation

Black birch	Betula nigra
Black oak	Quercus marilandica
Black alder	Alnus glutinosa
Black spruce	Picea mariana
Black poplar	Populus nigra



⇒ save as table, e.g. in SQL? Works well.

Putting quantities in relation



Das ist ein Auto	?	This is a car
Das ist ein Boot		This is a boat
Das Auto ist schön		This car is beautiful
Das Boot ist schön		This boat is beautiful
Ich mag Autos		I like cars

⇒ save as table, e. g. in SQL? Works well.

⇒ Gets very big very fast

⇒ No gain due to equal pre- and post-fixes

Putting quantities in relation

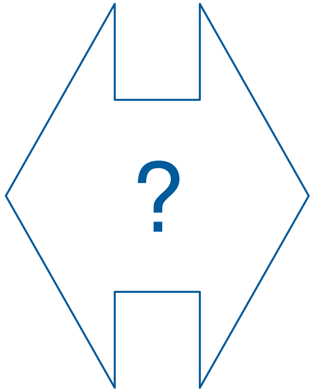
Das ist ein Auto

Das ist ein Boot

Das Auto ist schön

Das Boot ist schön

Ich mag Autos



This is a car

This is a boat

This car is beautiful

This boat is beautiful

I like cars

$\Rightarrow$ save as table, e. g. in SQL? Works well.

$\Rightarrow$ Gets very big very fast

$\Rightarrow$ No gain due to equal pre- and post-fixes

Putting quantities in relation

Das ist ein Auto

Das ist ein Boot

Das Auto ist schön

Das Boot ist schön

Ich mag Autos

Putting quantities in relation

Putting quantities in relation

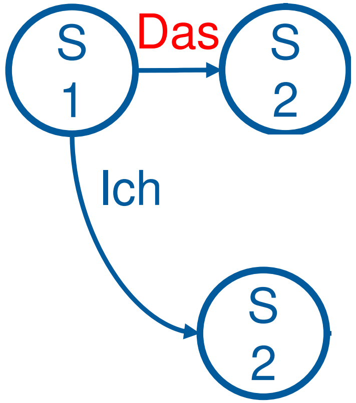
Das

Das

Das

Das

Ich



Putting quantities in relation

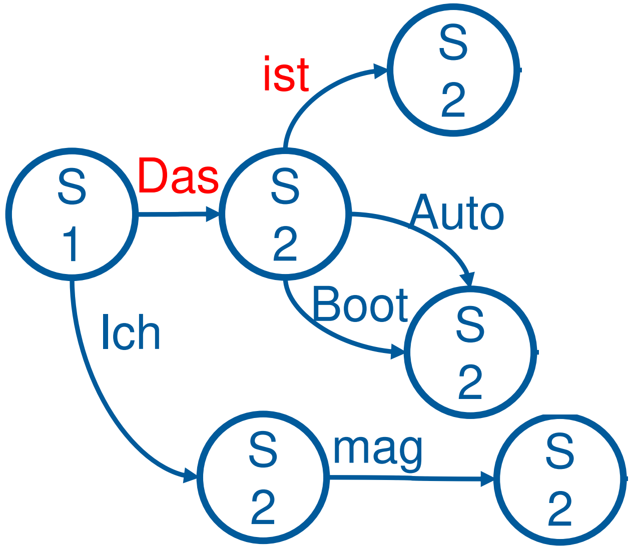
Das ist

Das ist

Das Auto

Das Boot

Ich mag



Putting quantities in relation

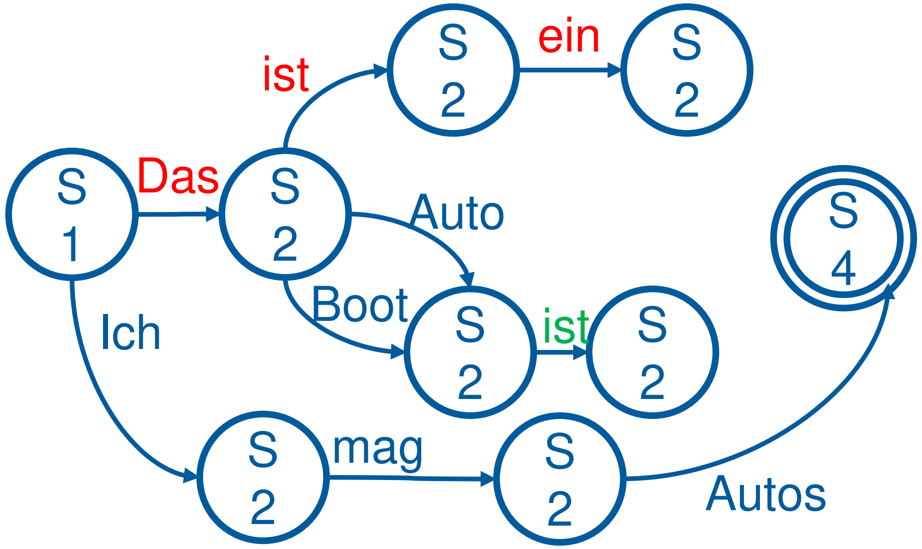
Das ist ein

Das ist ein

Das Auto ist

Das Boot ist

Ich mag Autos



Putting quantities in relation

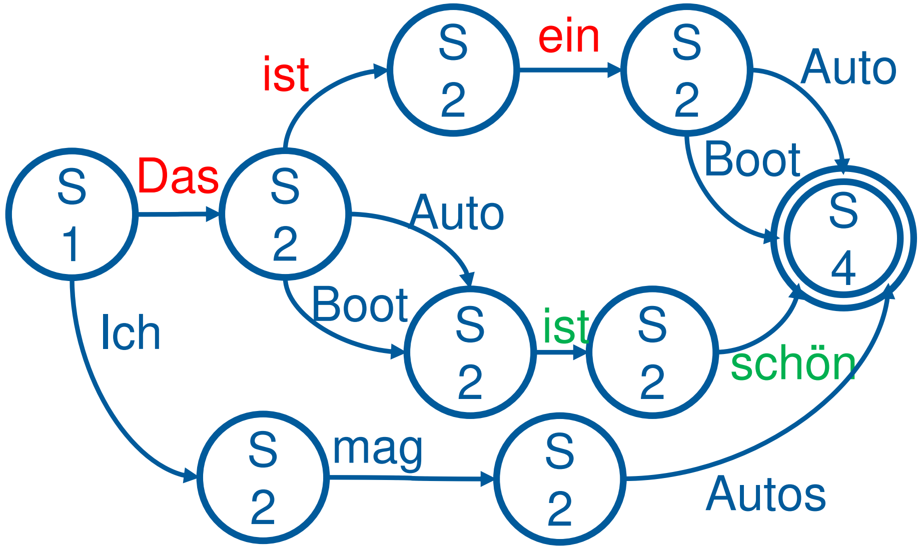
Das ist ein Auto

Das ist ein Boot

Das Auto ist schön

Das Boot ist schön

Ich mag Autos



Putting quantities in relation

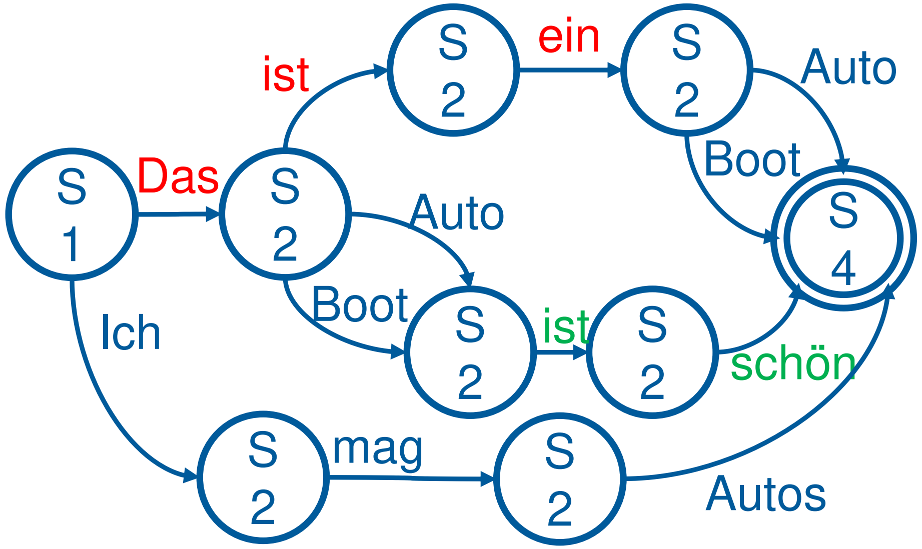
Das ist ein Auto

Das ist ein Boot

Das Auto ist schön

Das Boot ist schön

Ich mag Autos



An automaton constructed in this way is ALWAYS deterministic (but not minimal - unless you check the postfixes like we do).

Putting quantities in relation

This is a car

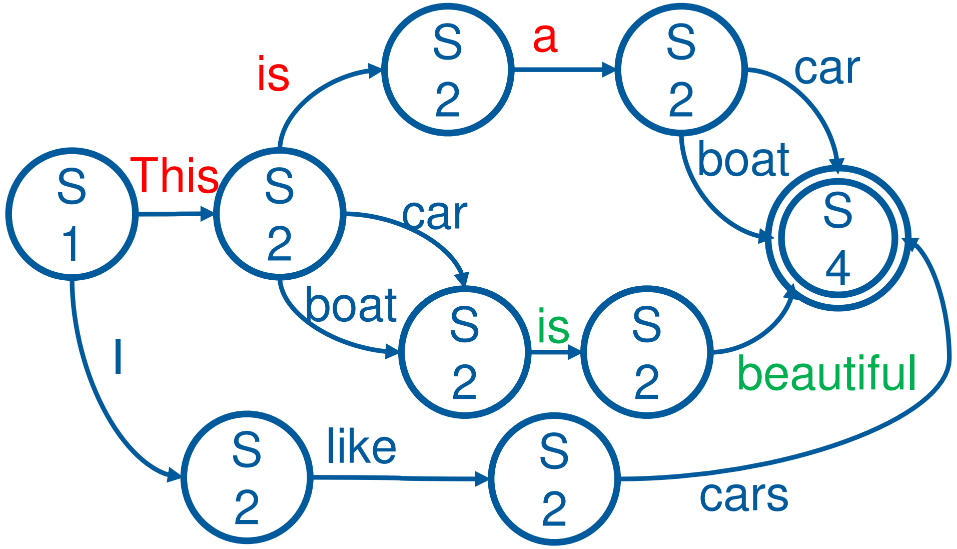
This is a boat

This car is beautiful

This boat is beautiful

I like cars

Putting quantities in relation



This is a car

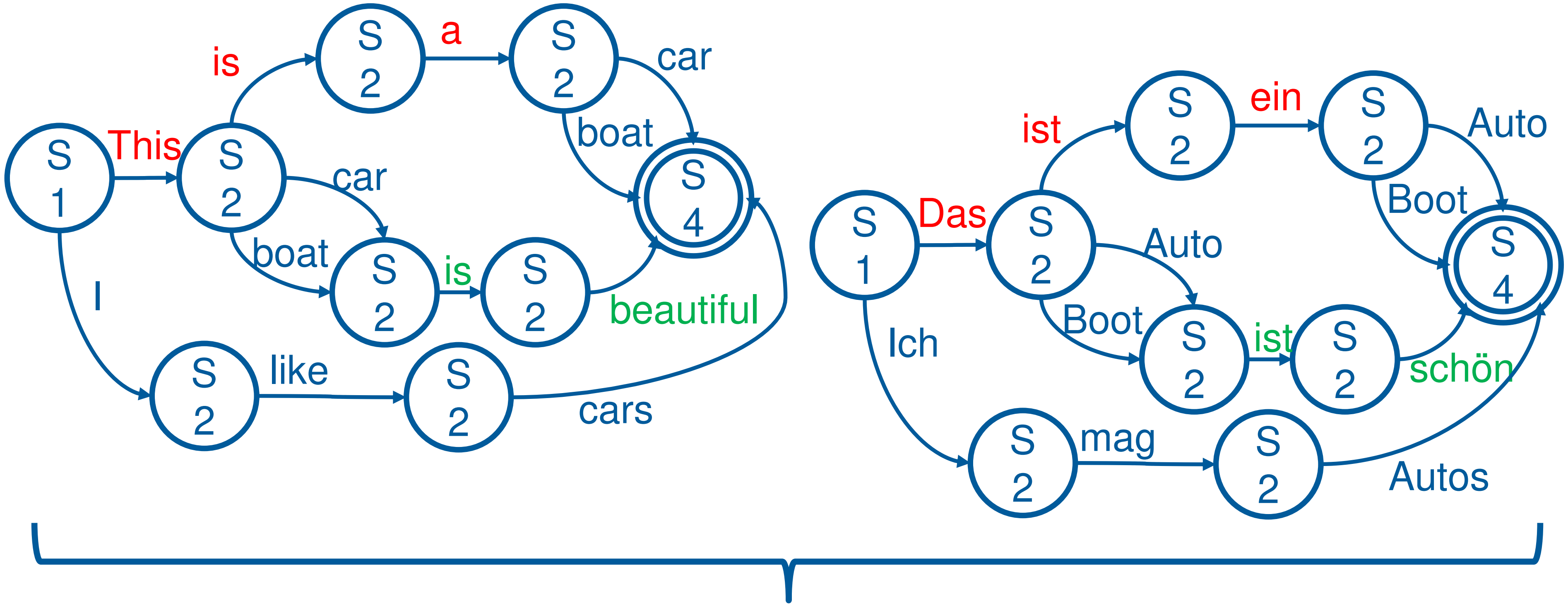
This is a boat

This car is beautiful

This boat is beautiful

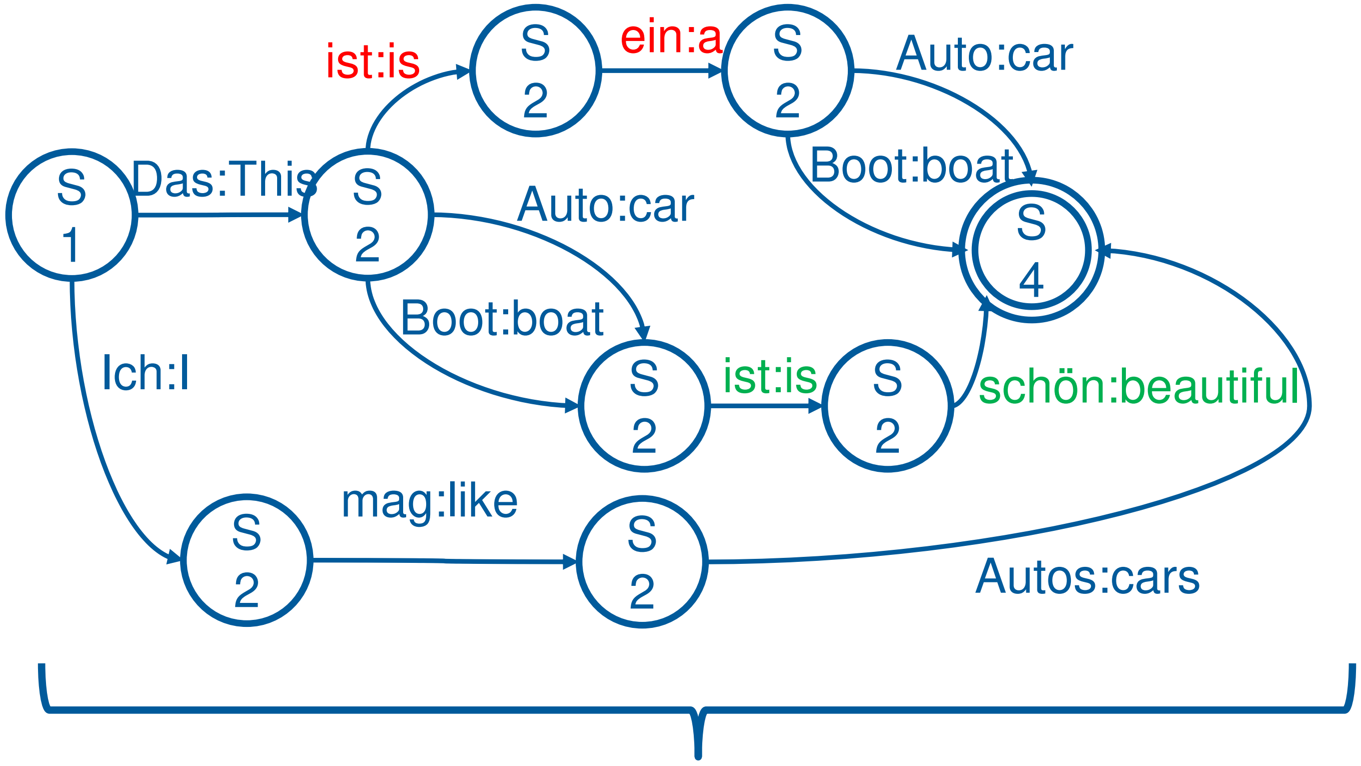
I like cars

Putting quantities in relation



Can we merge both automata?

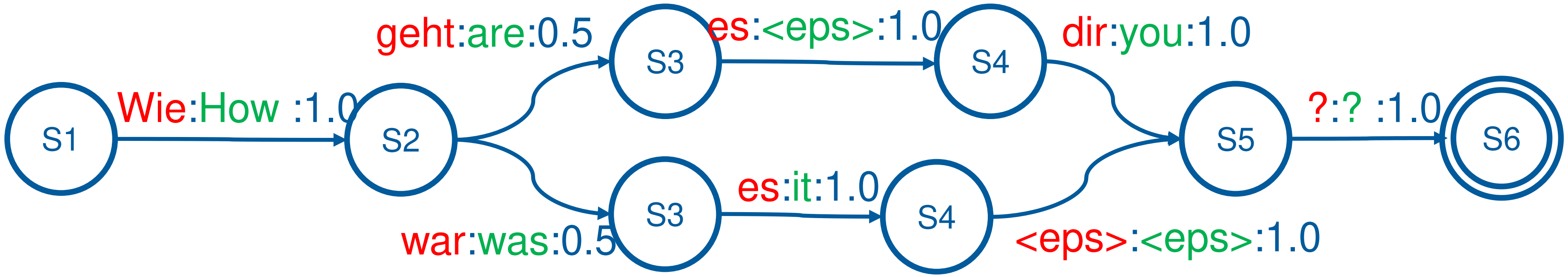
Putting quantities in relation



Finite State
Transducer

Weighted FSTs

- **Weighted Finite State Transducer (wFST)**



Transition defined by **<input>:<output>:<weight>**

Weight := depending on "semiring" a probability, log-likelihood. ...

https://en.wikipedia.org/wiki/Semiring (Algebraic structure)

Outlook

- **Composition of wFSTs (language modelling)**
- **Decoding of wFSTs (speech recognition)**

G2P: Grapheme-to-Phoneme

- **Maps the most likely phoneme sequence to grapheme sequence**
- **Structure is estimated by data**
- **Weights are estimated by data**

CAR	K AA1 R
CAR‘S	K AA1 R Z
CARA	K EH1 R AH0
CARA‘S	K EH1 R AH0 Z
CARA'VERAS	K AA2 R AH0 V EH1 R AH0 Z
CARABAJAL	K ER0 AE1 B AH0 JH AH0 L
CARABALLO	**K AE2 R AH0 B AE1 L OW0**
CARABALLO	**K AE2 R AH0 B EH1 L OW0**
CARACARA	K AA2 R AH0 K AA1 R AH0

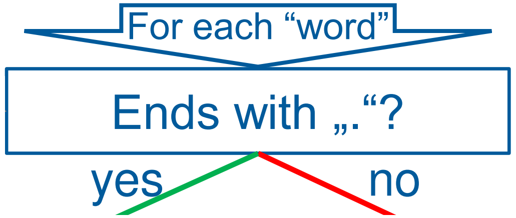
https://en.wikipedia.org/wiki/Phonological_rule

Decision Tree

A decision tree can be used to segment a text into separate sentences.

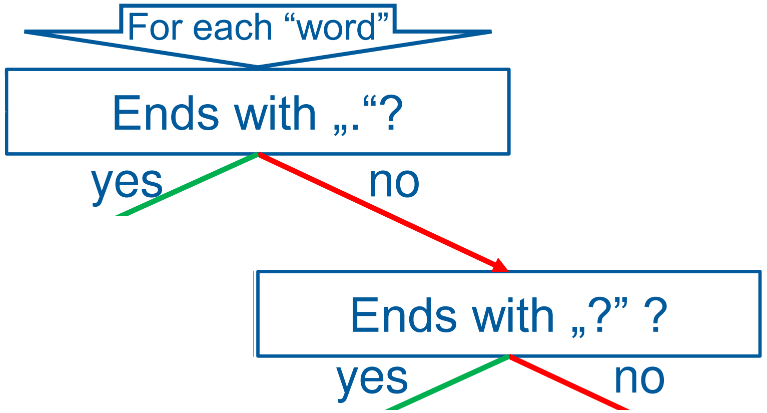
End-Of-Sentence Detection

`This is a text. A text is …`



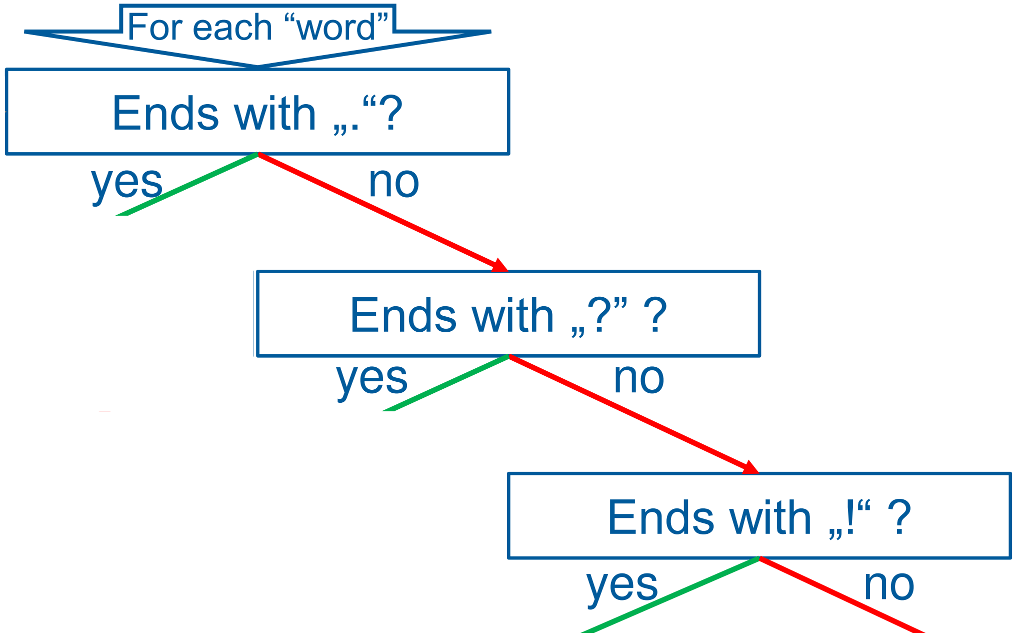
End-Of-Sentence Detection

```
This is a text. A text is …
```



End-Of-Sentence Detection

`This is a text. A text is …`



For each "word"

Ends with „."?

yes no

Ends with „?" ?

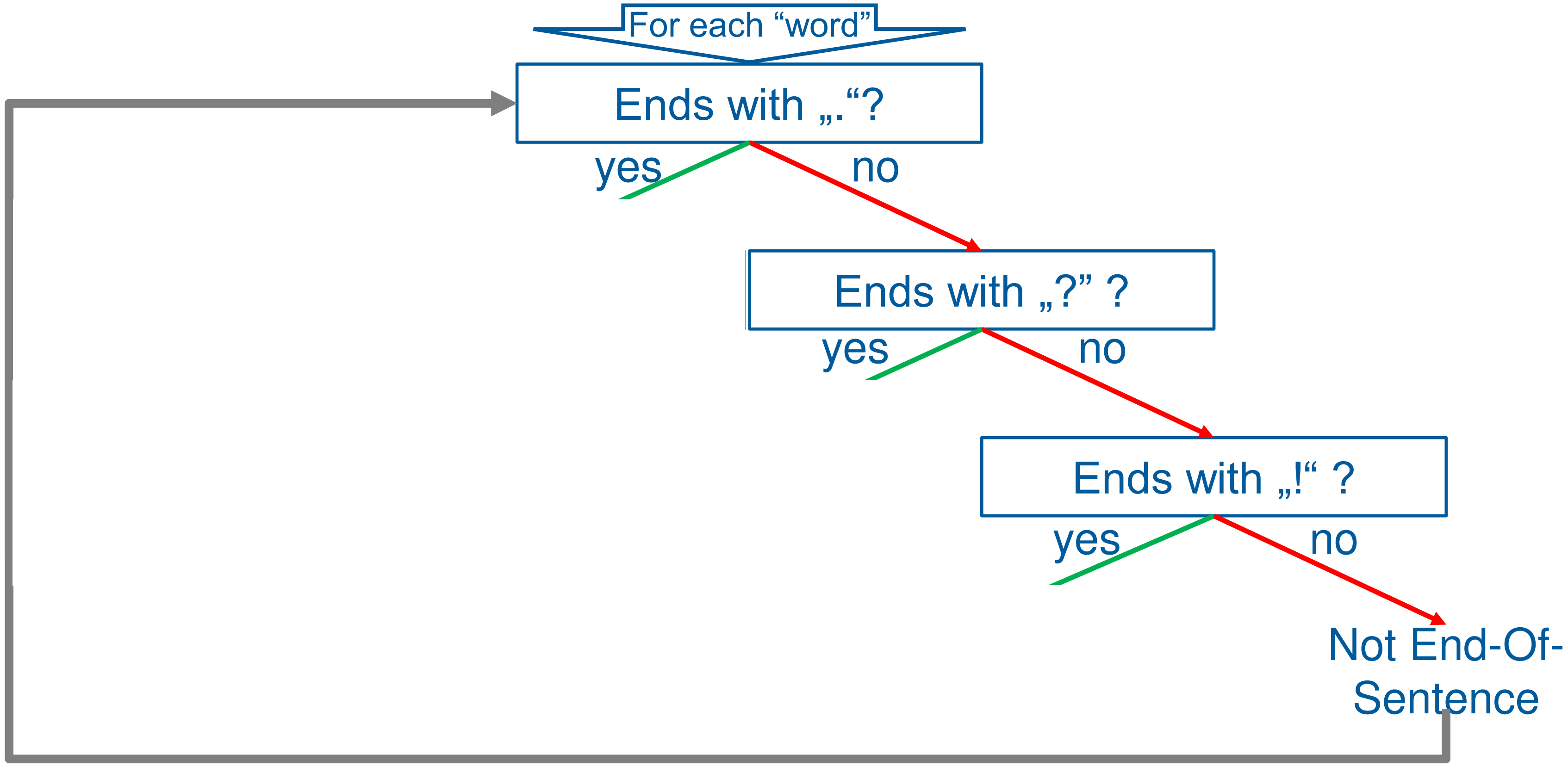
yes no

Ends with „!" ?

yes no

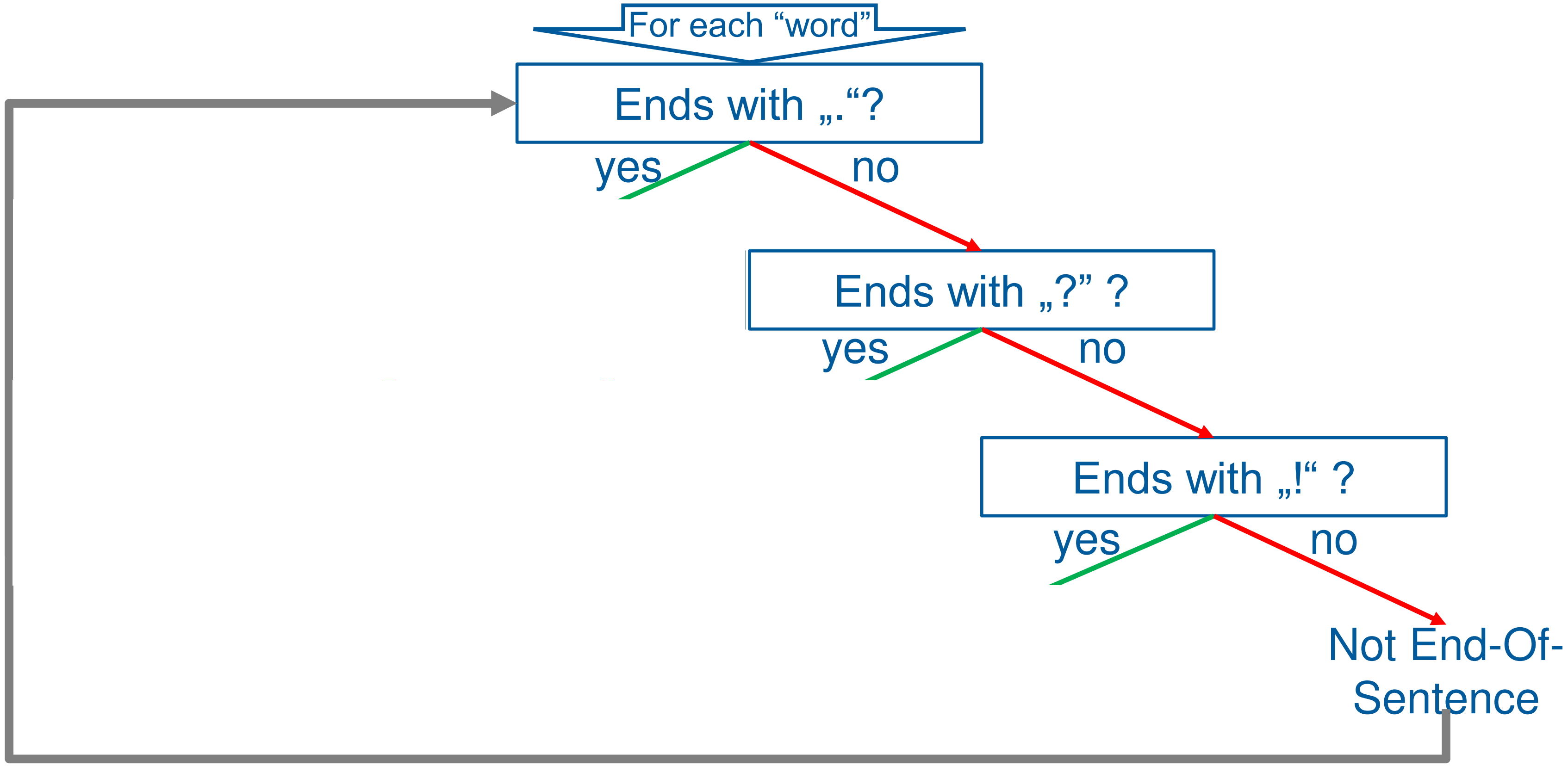
End-Of-Sentence Detection

`This is a text. A text is …`



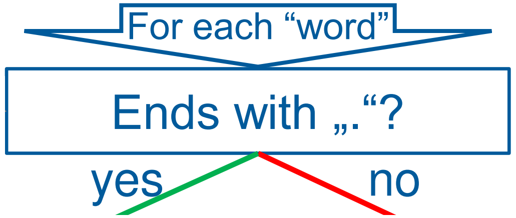
End-Of-Sentence Detection

`This is a text. A text is …`



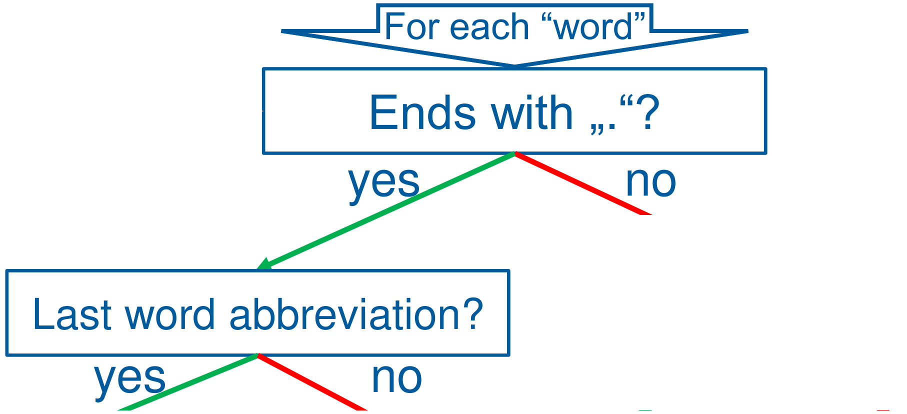
End-Of-Sentence Detection

`This is a text. A text is …`



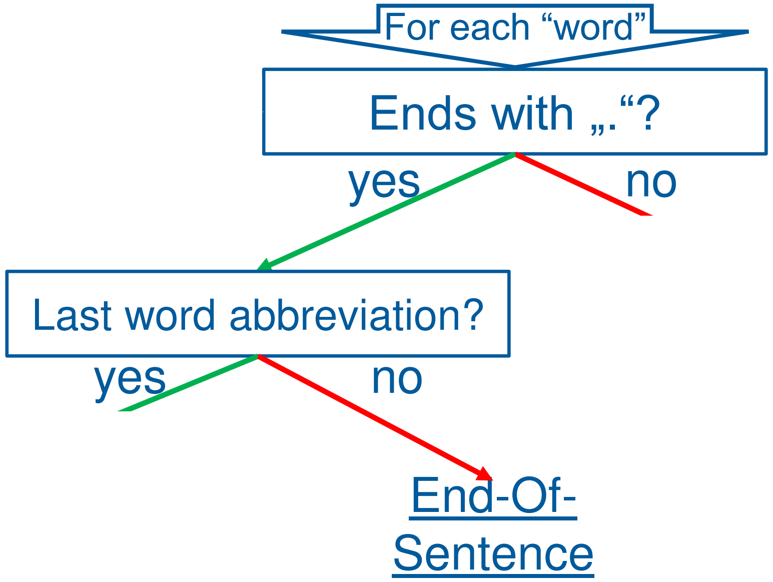
End-Of-Sentence Detection

`This is a text. A text is …`



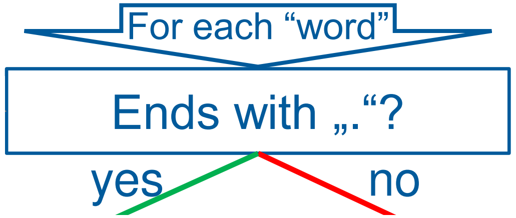
End-Of-Sentence Detection

`This is a text. A text is …`



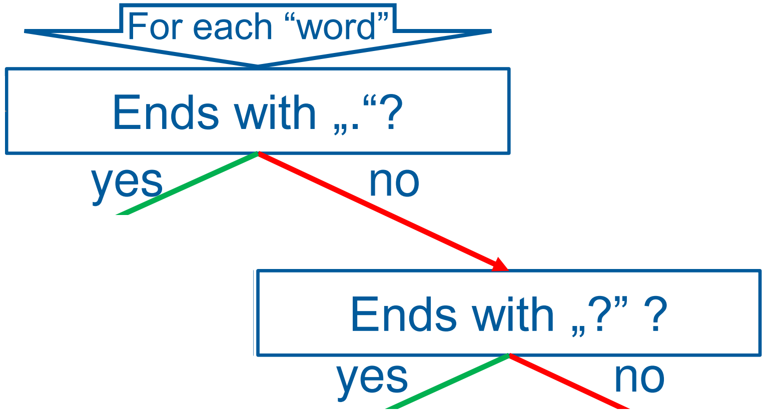
End-Of-Sentence Detection

`This is a text. A text is …`



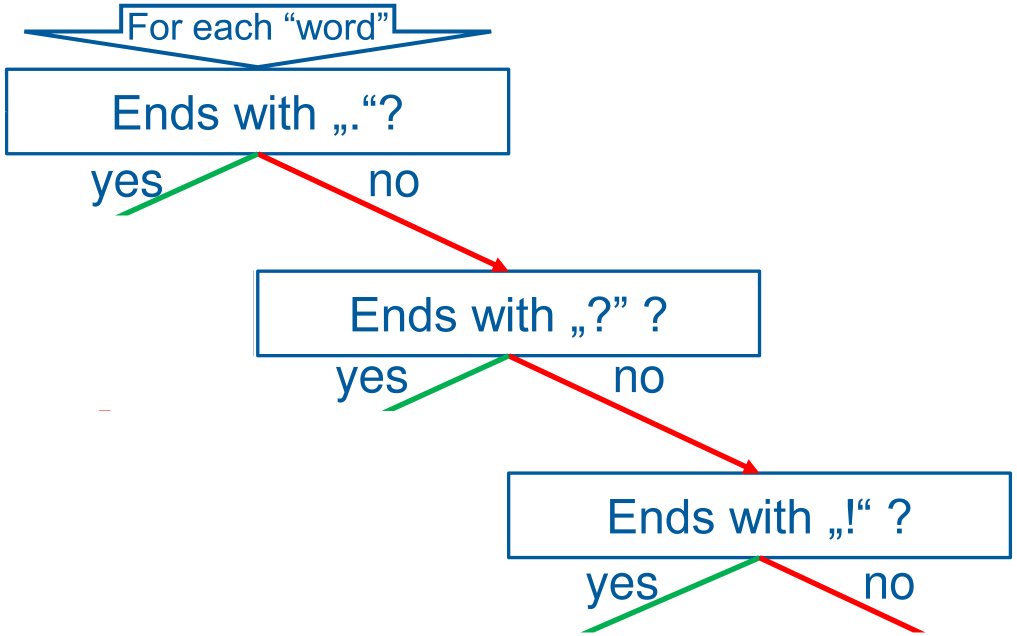
End-Of-Sentence Detection

```
This is a text. A text is …
```



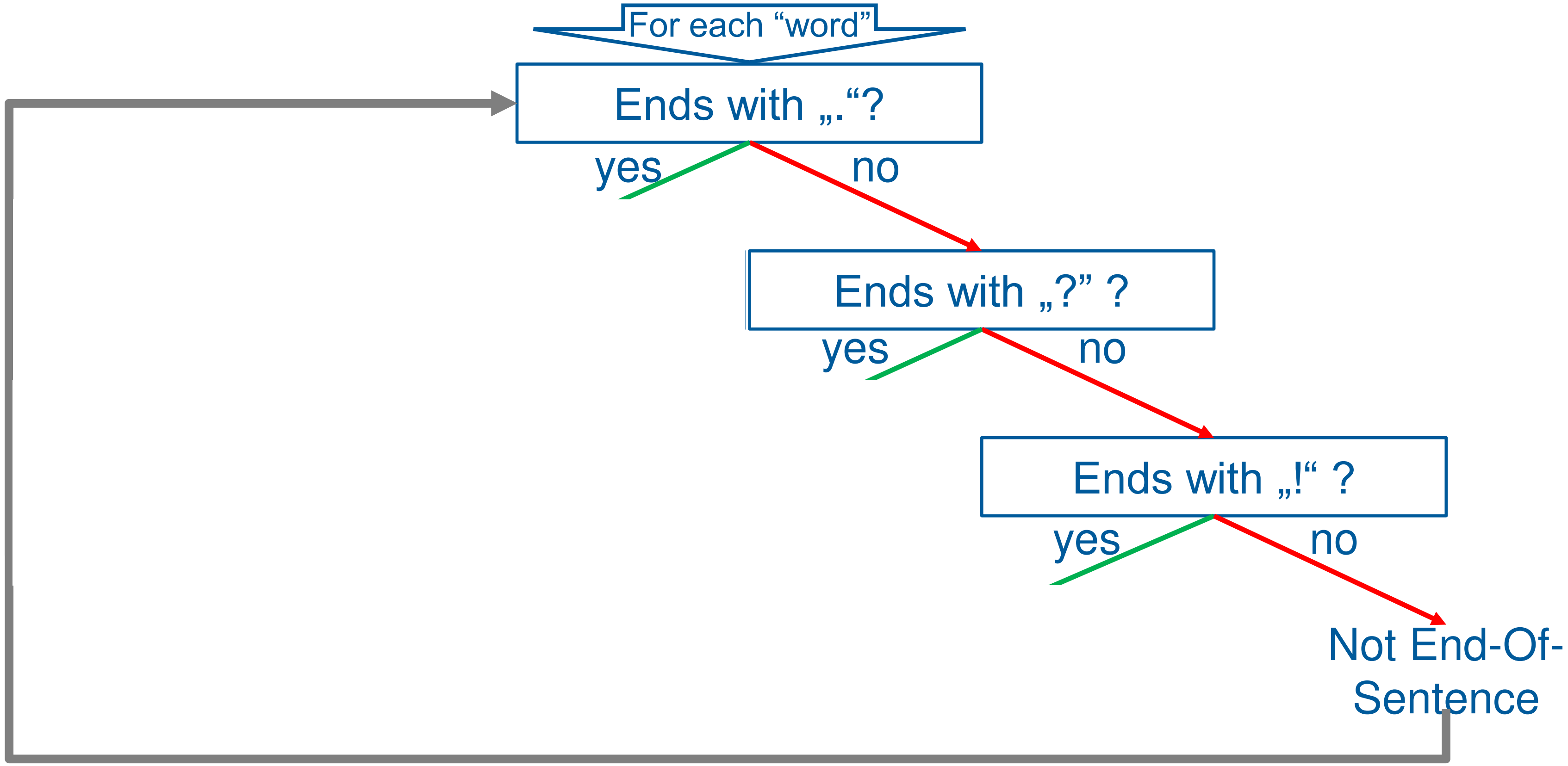
End-Of-Sentence Detection

`This is a text. A text is …`

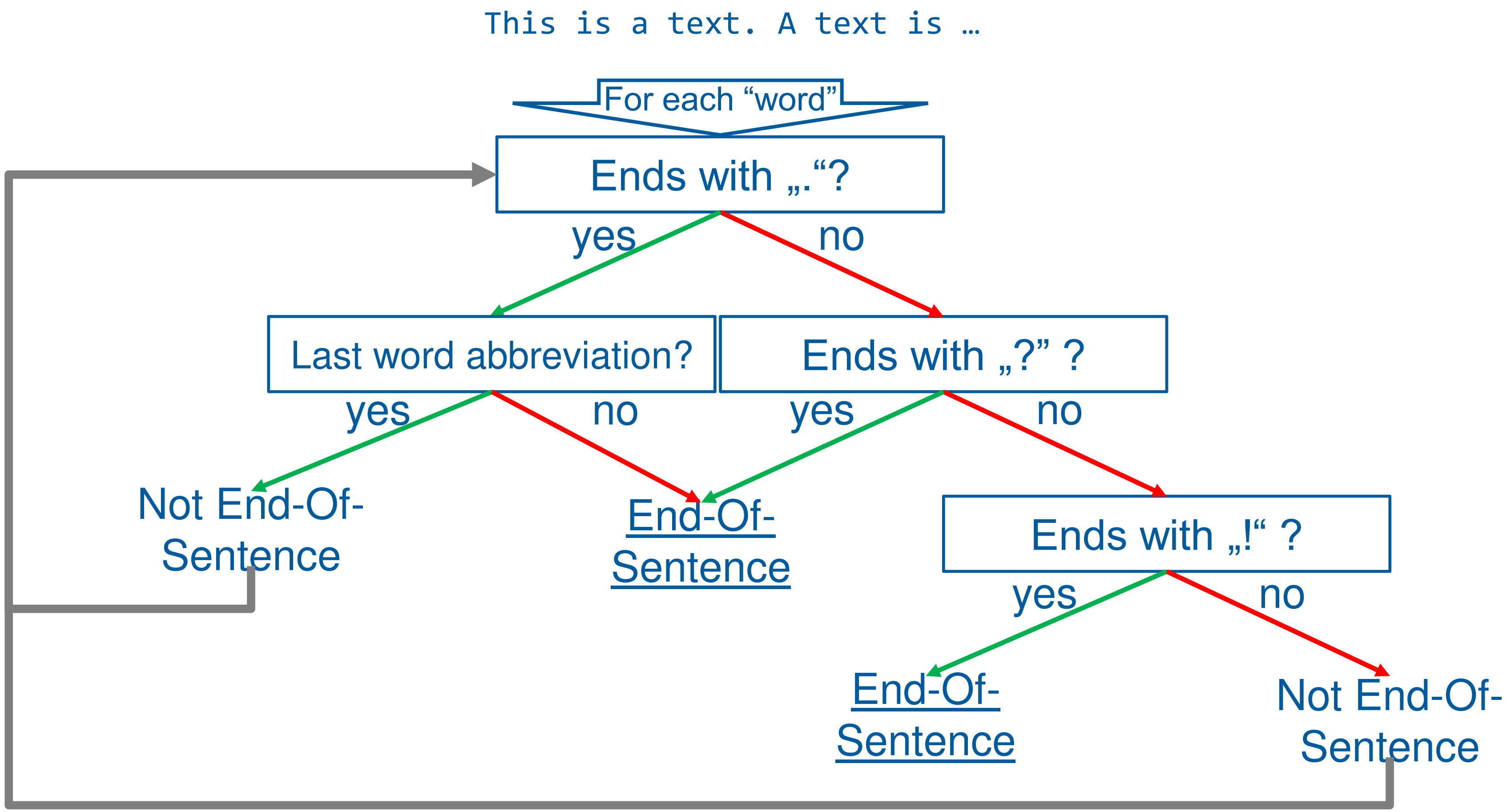


End-Of-Sentence Detection

`This is a text. A text is …`

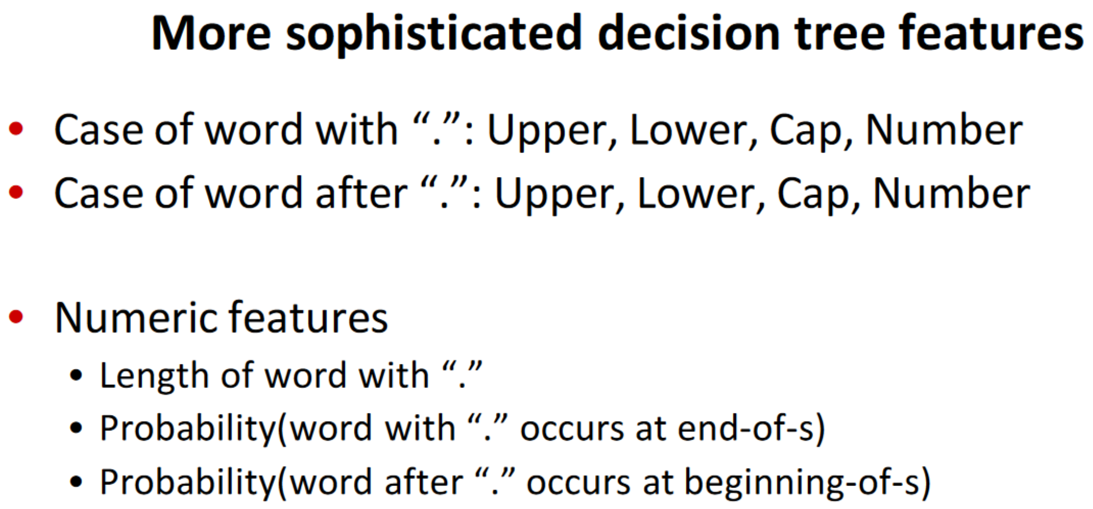


End-Of-Sentence Detection



Remarks

- **Easy to implement**
- **Finding adequate questions is challenging:**
  - Manually solvable only if there are not too many cases
  - Numberical constraints (word with > 10 characters)
  - Tree often generated via data analysis



More sophisticated decision tree features

- Case of word with “.”: Upper, Lower, Cap, Number
- Case of word after “.”: Upper, Lower, Cap, Number

- Numeric features
  - Length of word with “.”
  - Probability(word with “.” occurs at end-of-s)
  - Probability(word after “.” occurs at beginning-of-s)

Grammar

A grammar is a formal method to describe a (textual) language

Example Derivation of an English sentence

$$Sentence \Rightarrow Subject \quad Verb \quad Object$$

Example Derivation of an English sentence

$$\begin{aligned} \textit{Sentence} &\Rightarrow \textit{Subject} \quad \textit{Verb} \quad \textit{Object} \\ &\Rightarrow \textit{Noun-phrase} \quad \textit{Verb} \quad \textit{Object} \end{aligned}$$

Example Derivation of an English sentence

$$\begin{aligned}
\mathit{Sentence} &\Rightarrow \mathit{Subject} \quad \mathit{Verb} \quad \mathit{Object} \\
&\Rightarrow \mathit{Noun\text{-}phrase} \quad \mathit{Verb} \quad \mathit{Object} \\
&\Rightarrow \mathit{Article} \quad \mathit{Noun} \quad \mathit{Verb} \quad \mathit{Object}
\end{aligned}$$

Example Derivation of an English sentence

$$\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \mathit{Subject} & & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Noun\text{-}phrase} & & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object}
\end{array}$$

Example Derivation of an English sentence

$$
\begin{aligned}
\mathit{Sentence} &\Rightarrow \mathit{Subject}\quad \mathit{Verb}\quad \mathit{Object}\\
&\Rightarrow \mathit{Noun\text{-}phrase}\quad \mathit{Verb}\quad \mathit{Object}\\
&\Rightarrow \mathit{Article}\quad \mathit{Noun}\quad \mathit{Verb}\quad \mathit{Object}\\
&\Rightarrow \texttt{The}\quad \mathit{Noun}\quad \mathit{Verb}\quad \mathit{Object}\\
&\Rightarrow \texttt{The}\quad \texttt{students}\quad \mathit{Verb}\quad \mathit{Object}
\end{aligned}
$$

Example Derivation of an English sentence

$$\begin{array}{lllll}
\mathit{Sentence} & \Rightarrow & \mathit{Subject} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Noun\text{-}phrase} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} \quad \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} \quad \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} \quad \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} \quad \texttt{students} & \texttt{study} & \mathit{Object}
\end{array}$$

Example Derivation of an English sentence

$$
\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \multicolumn{2}{c}{\mathit{Subject}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \multicolumn{2}{c}{\mathit{Noun\text{-}phrase}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun\text{-}phrase}
\end{array}
$$

Example Derivation of an English sentence

$$
\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \multicolumn{2}{c}{\mathit{Subject}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \multicolumn{2}{c}{\mathit{Noun\text{-}phrase}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun\text{-}phrase} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun}
\end{array}
$$

Example Derivation of an English sentence

$$\begin{array}{llllll}
\textit{Sentence} & \Rightarrow & \multicolumn{2}{c}{\textit{Subject}} & \textit{Verb} & \textit{Object} \\
 & \Rightarrow & \multicolumn{2}{c}{\textit{Noun-phrase}} & \textit{Verb} & \textit{Object} \\
 & \Rightarrow & \textit{Article} & \textit{Noun} & \textit{Verb} & \textit{Object} \\
 & \Rightarrow & \texttt{The} & \textit{Noun} & \textit{Verb} & \textit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \textit{Verb} & \textit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \textit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \textit{Noun-phrase} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \textit{Noun} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \texttt{automata theory}
\end{array}$$

Motivation: Example Derivation of an English sentence

Sentence	⇒	*Subject*		*Verb*	*Object*
	⇒	*Noun-phrase*		*Verb*	*Object*
	⇒	*Article*	*Noun*	*Verb*	*Object*
	⇒	`The`	*Noun*	*Verb*	*Object*
	⇒	`The`	`students`	*Verb*	*Object*
	⇒	`The`	`students`	`study`	*Object*
	⇒	`The`	`students`	`study`	*Noun-phrase*
	⇒	`The`	`students`	`study`	*Noun*
	⇒	`The`	`students`	`study`	`automata theory`

Two types of words:

- *Subject, Verb, Noun*

Motivation: Example Derivation of an English sentence

$$
\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \multicolumn{2}{c}{\mathit{Subject}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \multicolumn{2}{c}{\mathit{Noun\text{-}phrase}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun\text{-}phrase} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \texttt{automata theory}
\end{array}
$$

Two types of words:

- *Subject, Verb, Noun* ⇔ **needs to be specified more**

Motivation: Example Derivation of an English sentence

$$
\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \multicolumn{2}{c}{\mathit{Subject}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \multicolumn{2}{c}{\mathit{Noun\text{-}phrase}} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun\text{-}phrase} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \texttt{automata theory}
\end{array}
$$

Two types of words:

- *Subject, Verb, Noun* ⇔ **needs to be specified more** ⇔ *Non-Terminal symbols*

Motivation: Example Derivation of an English sentence

Sentence	⇒	*Subject*		*Verb*	*Object*
	⇒	*Noun-phrase*		*Verb*	*Object*
	⇒	*Article*	*Noun*	*Verb*	*Object*
	⇒	The	*Noun*	*Verb*	*Object*
	⇒	The	students	*Verb*	*Object*
	⇒	The	students	study	*Object*
	⇒	The	students	study	*Noun-phrase*
	⇒	The	students	study	*Noun*
	⇒	The	students	study	automata theory

Two types of words:

- *Subject, Verb, Noun*	⇔	**needs to be specified more**	⇔	*Non-Terminal symbols*
- The, students, …	⇔	**don‘t need more explanation**	⇔	Terminal symbols

Motivation: Example Derivation of an English sentence

$$\begin{array}{llllll}
\mathit{Sentence} & \Rightarrow & \mathit{Subject} & & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Noun\text{-}phrase} & & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \mathit{Article} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \mathit{Noun} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \mathit{Verb} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Object} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun\text{-}phrase} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \mathit{Noun} \\
 & \Rightarrow & \texttt{The} & \texttt{students} & \texttt{study} & \texttt{automata theory}
\end{array}$$

Two types of words: *Non-Terminal symbols* `Terminal symbols`

A *Grammar* needs a set of rules that attribute information regarding non-terminal symbols

Definition and Example: Grammar

Finite set N **of non-terminals**

Finite set Σ of terminals: e.g. $\{a, b\}$

Start symbol: $\mathrm{S} \in N$

Set $\mathbf{P}$ **of production rules:** $\mathrm{P} = \{\mathrm{S} \Rightarrow a\mathrm{S}b, \mathrm{S} \Rightarrow ba\}$

Definition of a grammar

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aababb$$

Definition and Example: Grammar

Finite set N **of non-terminals**

Finite set Σ of terminals: e.g. $\{a, b\}$

Start symbol: $\mathrm{S} \in N$

Set P **of production rules:** $\mathrm{P} = \{\mathrm{S} \Rightarrow a\mathrm{S}b, \mathrm{S} \Rightarrow ba\}$

Definition of a grammar

Example derivation: $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aababb$

Language: $\{a^n bab^n \mid n \geq 0\} = \{ba, abab, aababb, aaababbb, \ldots\}$

Definition and Example: Grammar

Finite set N of non-terminals

Finite set Σ of terminals: e.g. $\{a, b\}$

Start symbol: $S \in N$

Set P of production rules: $P = \{S \Rightarrow aSb, S \Rightarrow ba\}$

Definition of a grammar

Example derivation: $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aababb$

Language: $\{a^n bab^n \mid n \geq 0\} = \{ba, abab, aababb, aaababbb, \ldots\}$

Language: „(a+b)=10“

Definition and Example: Grammar

Finite set N of non-terminals

Finite set Σ of terminals: e.g. $\{a, b\}$

Start symbol: $\mathrm{S} \in N$

Set P of production rules: $\mathrm{P} = \{\mathrm{S} \Rightarrow a\mathrm{S}b, \mathrm{S} \Rightarrow ba\}$

Definition of a grammar

Example derivation: $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aababb$

Language: $\{a^n bab^n \mid n \geq 0\} = \{ba, abab, aababb, aaababbb, \ldots\}$

Language: „(a+b)=10“

!not realizable with a DFA!

(Pushdown automaton [PDA] required)

Definition and Example: Grammar

Finite set N of non-terminals

Finite set Σ of terminals: e.g. $\{a, b\}$

Start symbol: $\mathrm{S} \in N$

Set P of production rules: $\mathrm{P} = \{\mathrm{S} \Rightarrow a\mathrm{S}b, \mathrm{S} \Rightarrow ba\}$

Definition of a grammar

Example derivation: $S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aababb$

Language: $\{a^n bab^n \mid n \geq 0\} = \{ba, abab, aababb, aaababbb, \ldots\}$

Language: „(a+b)=10“

Inspired by „Formal Languages and Automata Theory”, D. Goswami and K. V. Krishna, Nov. 5, 2010, Section 3
https://www.iitg.ac.in/dgoswami/Flat-Notes.pdf

Formal languages

Programming languages: C, Java, C#, JavaScript, …

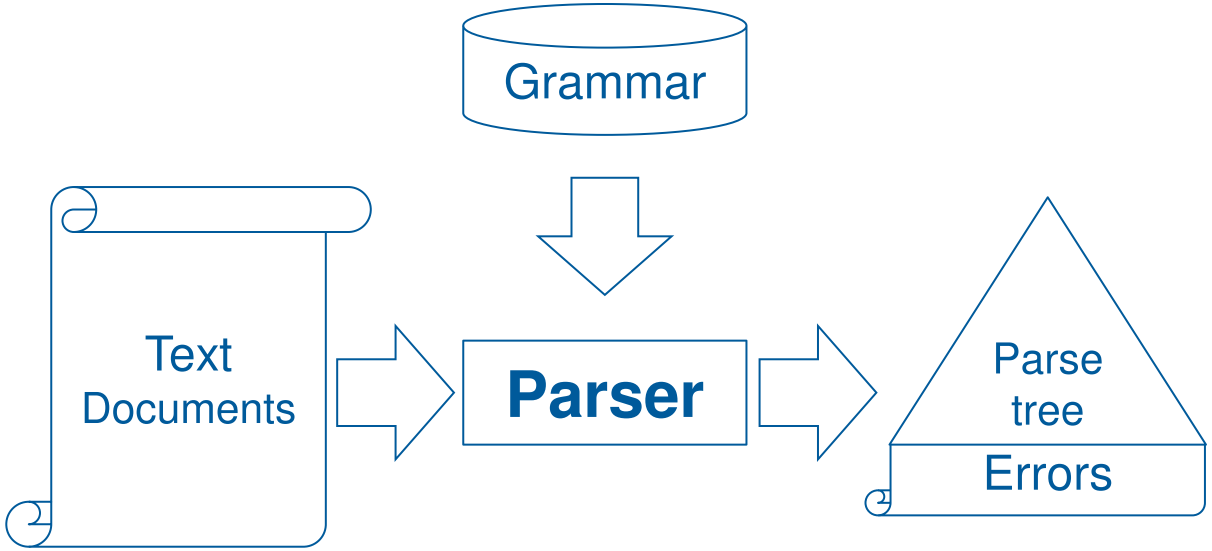
Domain-specific languages: TeX, BibTex, Mathematica, …

Data formats: log files, protocol data, …

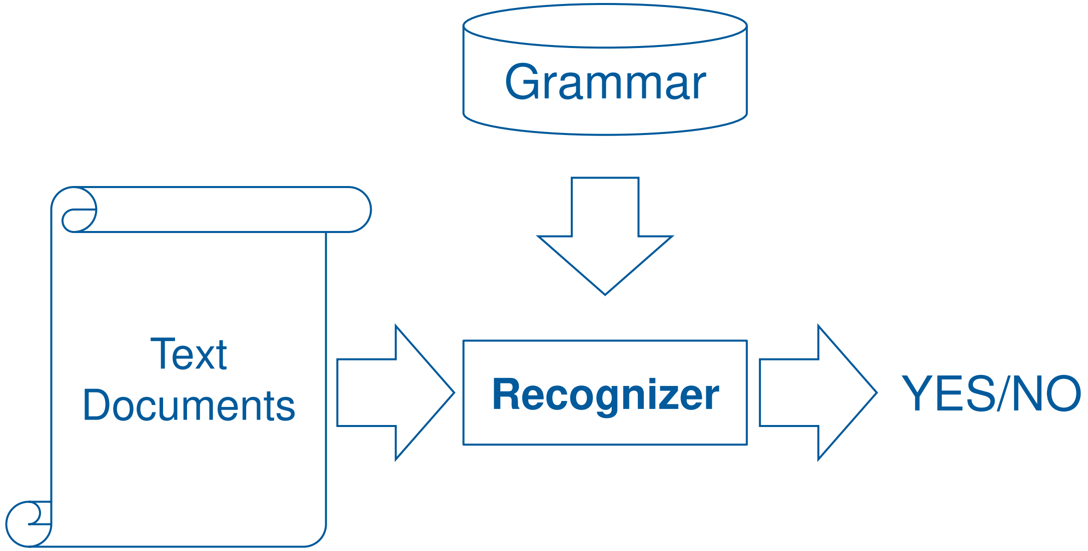
Natural Languages: ?

https://docs.python.org/3/reference/grammar.html

Parsing vs. Recognizing



Parsing vs. Recognizing



"Felix annoys the cat.“

https://hpi.de/friedrich/teaching/units/grammatiken.html

Text Generation

"Felix annoys the cat.“

"Felix" => "Stefan" : "Stefan annoys the cat."

"Felix annoys the cat.“

"Felix" => "Stefan" : "Stefan annoys the cat."

"annoys" => "catches" : "Stefan catches the cat.“

Text Generation

"Felix annoys the cat.“

"Felix" => "Stefan" : "Stefan annoys the cat."

"annoys" => "catches" : "Stefan catches the cat.“

"the" => "her" : "Stefan catches her cat.“

Text Generation

"Felix annoys the cat.“

"Felix" => "Stefan" : "Stefan annoys the cat."

"annoys" => "catches" : "Stefan catches the cat.“

"the" => "her" : "Stefan catches her cat.“

"cat" => "mouse" : "Stefan catches her mouse."

Text Generation

Der Hund beißt den Mann.

Den Mann beißt der Hund.

=> clear because of the declined articles.

The dog bites the man.

The man bites the dog.

=> Problem. This fact can be expressed in English only in one order.