Einführung

Katja Markert

Institut für Computerlinguistik
Uni Heidelberg
markert@cl.uni-heidelberg.de
mit Folien von Yannick Versley und Anette Frank

October 24, 2019

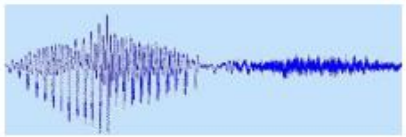
Übersicht

Übersicht

Allgemeine Definition

- Computerlinguistik beschäftigt sich mit der maschinellen Verarbeitung natürlicher Sprache.
- Strukturelle Eigenschaften und Verarbeitungsmechanismen natürlicher Sprache.

akustisches Signal (speech)
geschriebener Text (text)



Johanna liest ein Buch.

Bedeutung (logische Form)
Kognitiver Inhalt

Johanna(x)
buch(y)
lesen(e,x,y)

„Buch“

buch(…)

CL Ziele

- Studium der formalen Eigenschaften von Sprache: Wie funktioniert Sprache? Wie kann ich dies formal repräsentieren und modellieren?

  Beispielfrage: Welche Ausdrücke sind korrektes Deutsch/Englisch?
- Erklärung / Simulation von Sprachverständnis
  - Analyse
  - Generierung
  - Übersetzung

  Beispielfrage: Was sagen Menschen? Welche Muster im Sprachgebrauch gibt es?

Übersicht

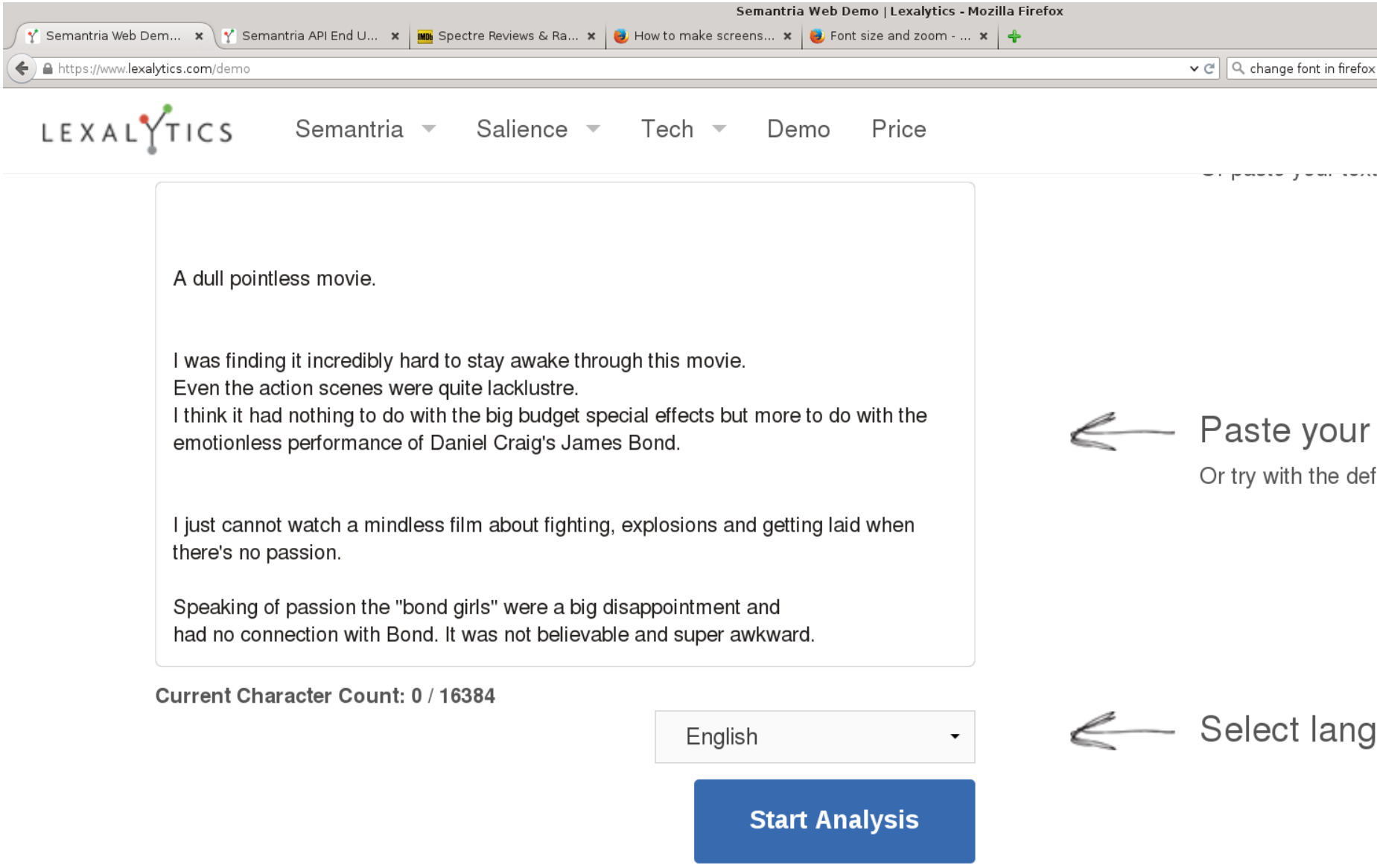
Beispiel I: Opinion Mining

- Extraktion von Meinungen und Emotionen aus Texten
- Typisch: Vorhersage von Verkäufen oder Filmerfolgen oder Verfolgung von Markenreputation
- Social Media Monitoring

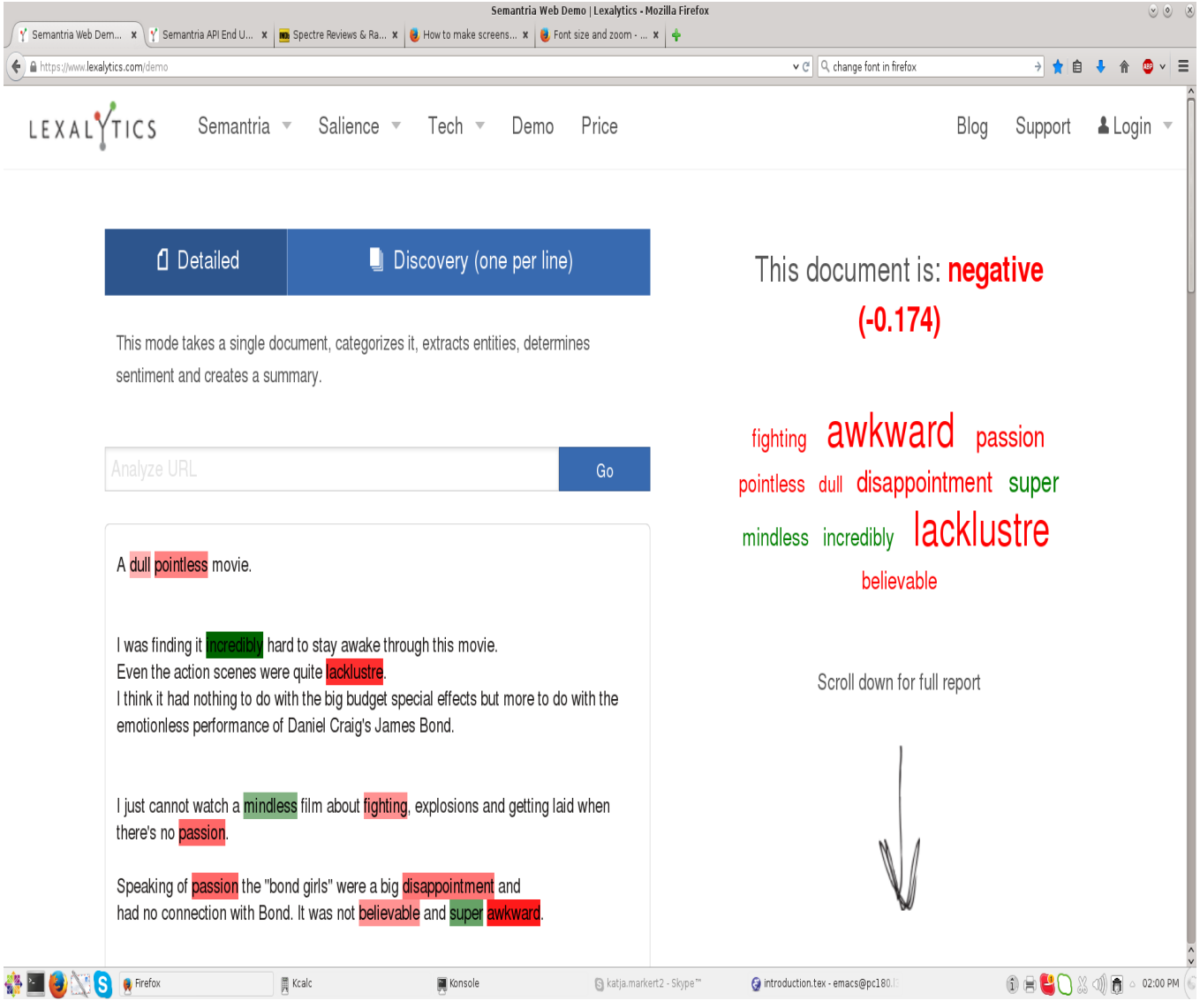
Demos:

- Lexalytics Demo `https://www.lexalytics.com/demo` (braucht account)
- Sentiment Treebank Demo `http://nlp.stanford.edu:8080/sentiment/rntnDemo.html` (auch Teil der CoreNLP pipeline `https://stanfordnlp.github.io/CoreNLP/index.html`)

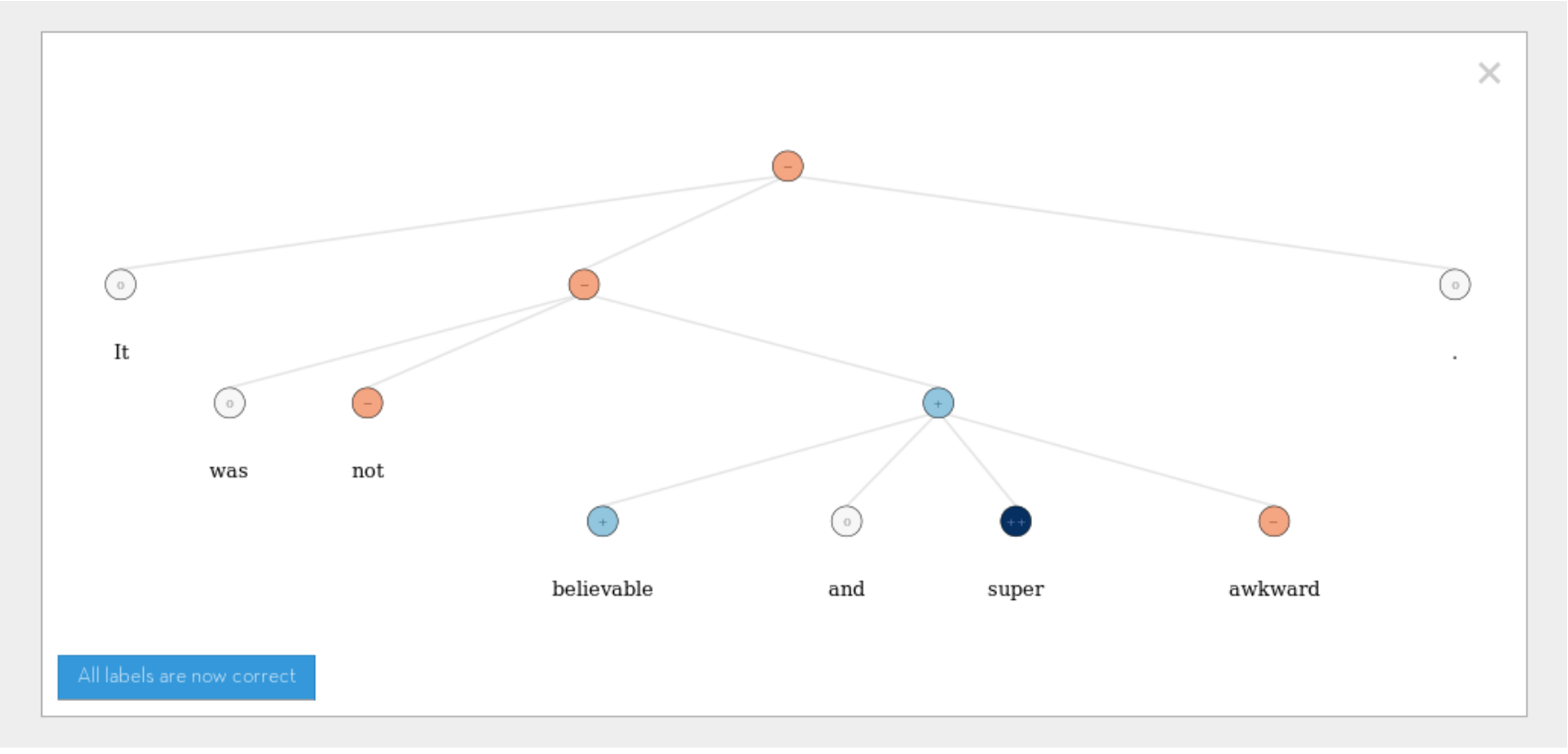
Lexalytics Input



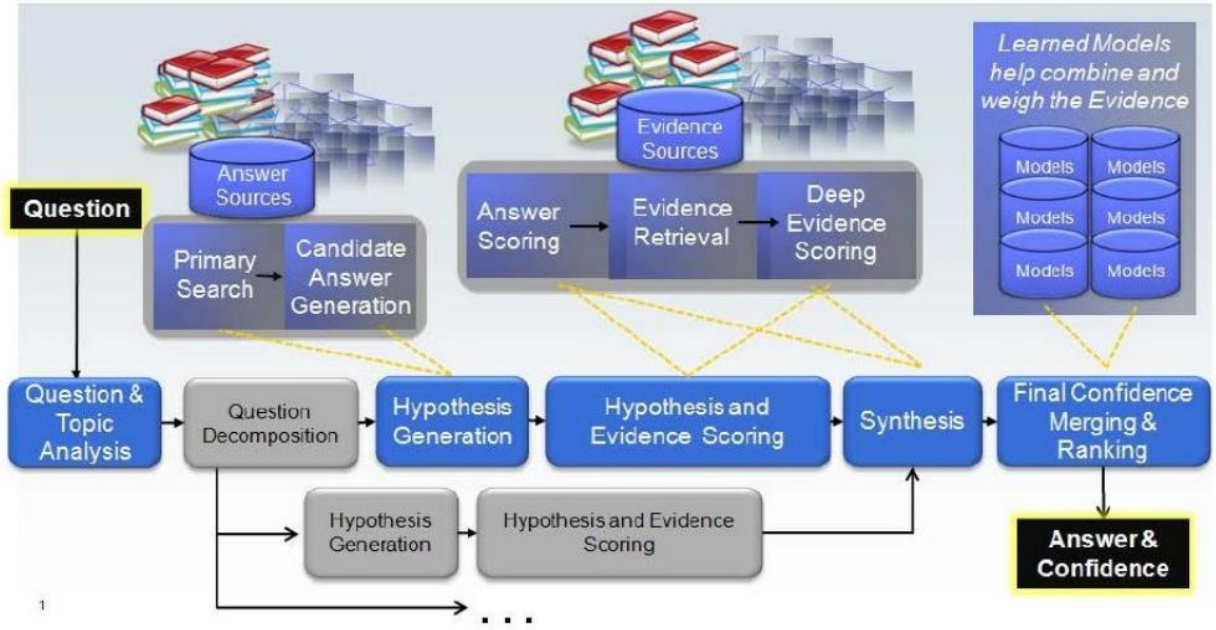
Lexalytics Output



Stanford sentiment treebank



Beispiel II: Question Answering



- Suche nach Antworten auf spezielle Fragen
  - Faktoid: *Wann wurde Leonardo da Vinci geboren?*
  - Komplex: *Wie alt war Leonardo da Vinci als Michelangelo geboren wurde?*
  - Schlussfolgerung*: Leonardo da Vinci war älter als Michelangelo.*
- Linguistische Analyse der Frage: Fragetyp, Thema
- Suche nach passender Information (Kandidaten), Evaluation von Hypothesen
- Aufbereitung der Antwort

Beispiel II: Question Answering

Demo: Wolfram Alpha `http://www.wolframalpha.com/` (accessed 14.10.2019)

- When was Leonardo da Vinci born? $\longrightarrow$ Saturday, April 15, 1452
- How old was Leonardo da Vinci when Michelangelo was born? $\longrightarrow$ Result: 22 years 10 months 18 days
- Has Elvis died $\longrightarrow$ Yes
- Has Elvis kicked the bucket? $\longrightarrow$ Yes
- Has Elvis given up the ghost? $\longrightarrow$ Interpreting as “color ghost”.
- When did Shakespeare die? $\longrightarrow$ Tuesday, April 23, 1616
- Shakepeare wrote many plays. When did he die? $\longrightarrow$ Input interpretation “wrote” (English word).

Multilinguale Anwendungen: Maschinelle Übersetzung

- Sicher eine der bekanntesten und wichtigsten Anwendungen
- Unterschiedliche linguistische Eigenschaften verschiedener Sprachen
- Interpretation, Wissen und Übersetzung
- Varianten: Vollübersetzung (wissensbasiert beispielbasiert statistisch (SMT)), Unterstützte Systeme (HAMT) human-aided MT, Unterstützende Systeme (MAHT) machine-aided HT

MT: Vollautomatisch

Google Translate (14.10.2019)

- Im Wesentlichen handelt Star Wars vom ständig andauernden Kampf zwischen Gut und Bse. **Essentially, Star Wars is about constant struggle between good and evil.**
- Die Kaffeemaschine ist kaputt. Ich lasse sie reparieren. **The coffee maker is broken. I have her repaired.**

DeepL (14.10.2019)

- Die Kaffeemaschine ist kaputt. Ich lasse sie reparieren. **The coffee machine is broken. I have it repaired.**
- The coffee machine is broken. I will have it repaired. **Die Kaffeemaschine ist kaputt. Ich werde es reparieren lassen.**

Übersicht

Beispielfragen I

1. Welchem deutschen Wort entspricht
   [raːt]
2. Wie wird das Wort *rasten* ausgesprochen?
3. Wie verstehen Sie die folgende Sprachsequenz:
   [hasm moˈmɛnˈtsaɪt]

Beschreibungsebene: Phonetik und Phonologie

- Phonetik und Phonologie
  - Artikulatorische Merkmale und Lautstruktur
  - Wortsegmentierung, Aussprache, Prosodie
- **Phonem:** Kleinste Spracheinheit, die Bedeutungsunterschiede ausmacht
- **Homophone:** verschiedene Worte mit gleicher Aussprache
- Variation in Aussprache
  - **lexikalisch**: durch Ambiguität; oft soziolinguistisch
  - **allophonisch**: meist kontextuelle Variation

Spracherkennung und Sprachsynthese

Spracherkennung: Übersetzung von gesprochener Sprache in Text

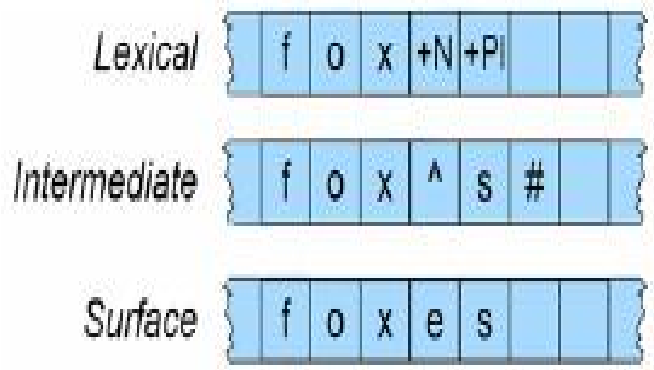
Sprachsynthese/Text-to-Speech: künstliche Erzeugung der menschlichen Sprechstimme

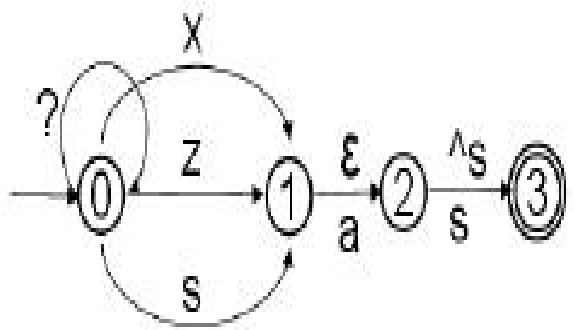
Beispielfragen II

1. Ist *riche* ein deutsches Wort? Wie ist es mit *freche*?
2. Is *un* ein Wort? Hat es eine Bedeutung?
3. Wie viele “Bedeutungseinheiten” hat *Frechheiten*? Wie ist es mit *Staubecken*?

Beschreibungsebene: Morphologie

- **Morphologie:** Beschreibung von Bildung und Struktur von Wörtern
- **Morpheme:** kleinste bedeutungstragende Einheit
- Systematische Beziehungen zwischen Wörtern und Wortform: Flexion, Derivation, Komposition (*frech - Frechheit - Frechdachs)*
- Prozesse/Regeln zur Erzeugung von Wortformen
- **Morphologisches Parsing:** Finde die Morpheme eines Wortes $\longrightarrow$ endliche Automaten, formale Sprachen



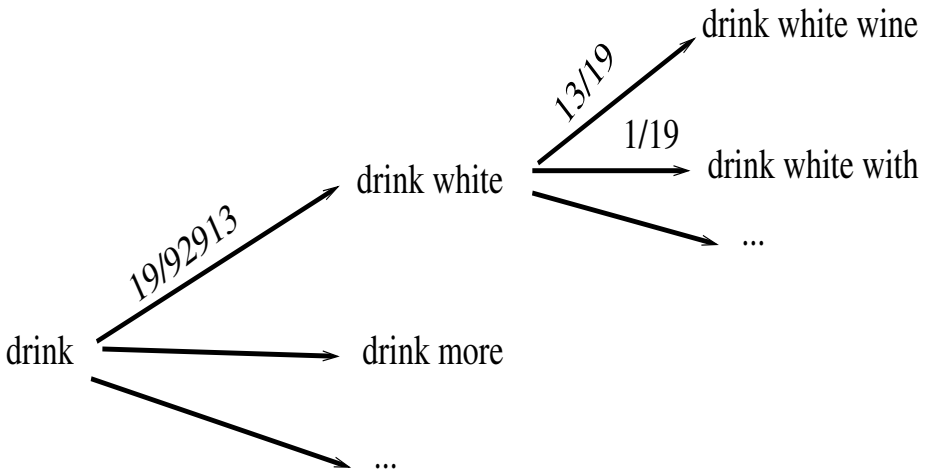


Beispielfragen III

1. Welche Worte folgen wahrscheinlich der (englischen) Sequenz “white ...” ?
2. Welche Worte folgen wahrscheinlich der englischen Sequenz “drink white ...” ?

Ngram Modelling

ngram modelling: Ordne Wortsequenzen Wahrscheinlichkeiten zu
$\longrightarrow$ Wahrscheinlichkeitstheorie und Schätzung

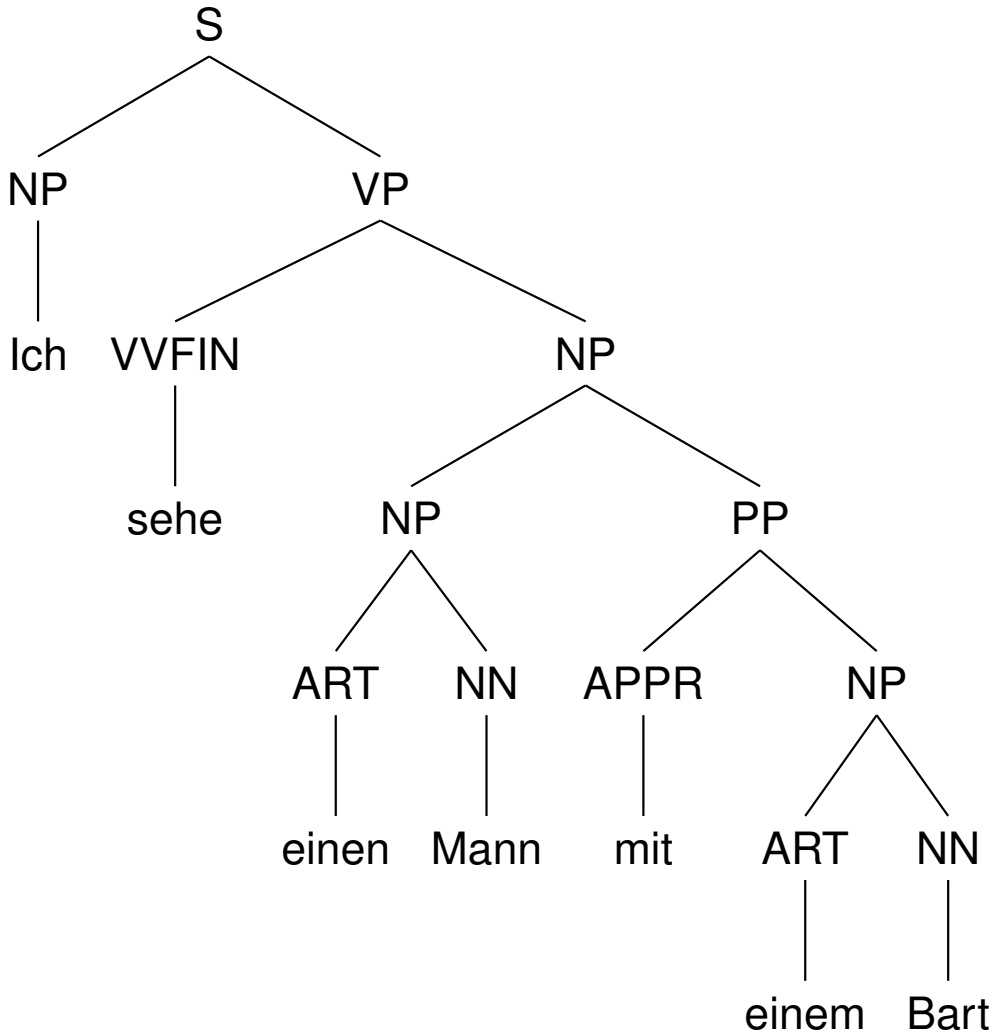


$$P(w_i|w_1 \dots w_{i-1}) = \frac{P(w_1 \dots w_i)}{P(w_1 \dots w_{i-1})} \approx \frac{f(w_1 \dots w_i)}{f(w_1 \dots w_{i-1})}$$

Alternativ: ngram modelling mit neuronalen Netzen

Beispielfragen IV

Ist “Ich sehe einen Mann mit meinem Fernglas” ein **grammatikalisch richtiger Satz**? Hat der Mann mein Fernglas oder sehe ich ihn? Wie steht es mit “Ich sehe einen Mann mit einem Bart”?



Beschreibungsebene: Syntax

- **Syntax:** beschreibt strukturelle Beziehung zwischen Wörtern
- Typische Fragen: Grammatikalität?
- **Syntaktische Regeln**: Prozesse, die Sätze generieren können

  VP $\rightarrow$ VVFIN NP

  VP $\rightarrow$ VP PP
- **Parsing:** Zähle die syntaktischen Strukturen (parses) von Sätzen auf und entscheide Dich für eine präferierte Struktur; benutzt syntaktische Regeln
- Präferenzen: oft auch wieder statistisch

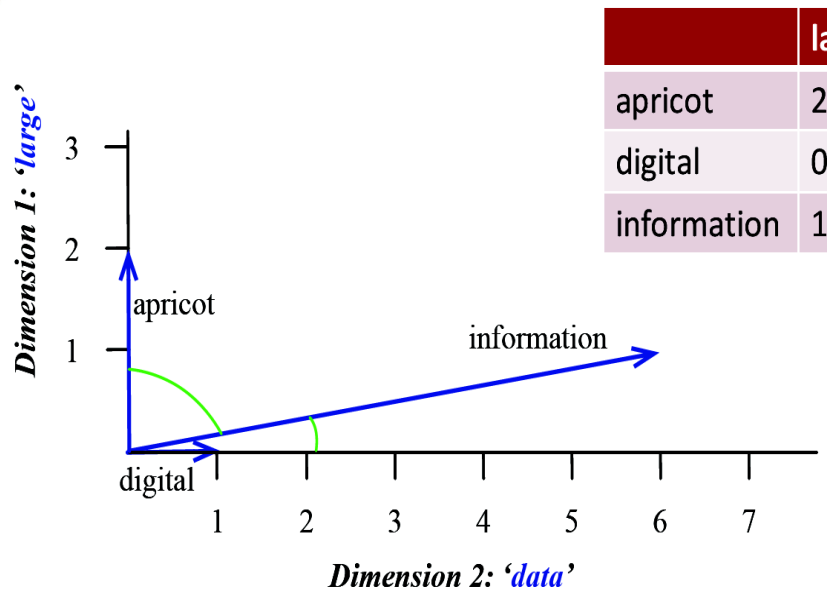
VP	$\rightarrow$	VVFIN NP	0.7
VP	$\rightarrow$	VP PP	0.3

Beispielfragen V

1. Kann *grün* in *ein grüner Junge* durch *unerfahren* ersetzt werden? Wie steht es bei *ein grüner Baum*? Wieviele Bedeutungen von *grün* gibt es?
2. Welches Wort fällt Ihnen als erstes ein, wenn Sie hören: *Apfel*
3. Was bedeutet *Jeder Holländer besitzt einen Wohnwagen*?

Bereiche: Semantik

- **Semantik:** Wissenschaft der Bedeutung
- **Lexikalische Semantik:** Wortbedeutung
- Distributionelle Semantik: Worte, die in gleichen Kontexten auftauchen teilen, semantische Bedeutung
- Vektorsemantik, lineare Algebra, neuronale Netze



	large	data
apricot	2	0
digital	0	1
information	1	6

Bild von Dan Jurafsky, Stanford University

Bereiche: Semantik

- **Satzsemantik und das Kompositionalitätsprinzip:** Bedeutung eines Satzes ergibt sich aus der Bedeutung seiner Teile. Ausnahmen?
- **Semantische Analyse:** Bilde linguistischen Input auf formale Bedeutungsrepräsentationen ab. Semantische Relationen.

$$\forall x\ (Holländer(x) \rightarrow \exists y\ wohnwagen(y) \wedge besitzen(x,y))$$

- Prädikatenlogik, Lambda-Kalkül (siehe *Einführung in die Logik*)

Beispielfragen VI

1. *Shakespeare war einer der produktivsten Schrifsteller des Elisabethanischen Zeitalters und unter König James immer noch beliebt.* ***Er*** *schrieb 38 Stücke.* Auf wen referiert *er*?.
2. Wie sind die folgenden Satzteile relationiert? *Anna arbeitete an ihren ECL Übungen während sie Musik hörte.*
3. Wie sind die folgenden Satzteile relationiert? *Anna arbeitete an ihren ECL Übungen während Mia auf eine Party ging.*

Beschreibungsebene: Diskurs

- **Diskurs:** Linguistik über die Satzgrenze hinaus.
- Hauptfrage: was macht einen Text kohärent?
- **Referenzresolution:** Welche Entitäten in einem Text sind koreferent?
- **Rhetorische Relationen/Diskursrelationen:** Diskursrelationen zwischen Sätzen (z.B. temporal, Kausal etc).

Ambiguität: ein abschreckendes Beispiel

Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.

Wieviele Lesarten besitzt dieser Satz?

$2 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 4 \cdot 2 \cdot 2 \cdot 7 \cdot 2 = 258.048$

Quelle: Hans Uszkoreit

Ambiguität: ein abschreckendes Beispiel

Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.

- *Früher* kann eigenständiges Adverb oder Komparativ von früh sein (2);
- die Verbform *stellten* ist ambig zwischen Präteritum und Konjunktiv (2)
- die Nominalphrase *die Frauen* kann Subjekt oder Objekt des Satzes sein (2)
- *am Wochenende* kann *die Insel*, *die Frauen* oder das Verb modifizieren (3);
- *mit Blumenmotiven* kann sich auf *die Kopftcher* beziehen, ein Instrument der Herstellung sein oder ein Adjunkt im Sinne von gemeinsam mit Blumenmotiven(3);
- *Her* hat auch eine direktionale Bedeutung (2)

Quelle: Hans Uszkoreit

Ambiguität: ein abschreckendes Beispiel

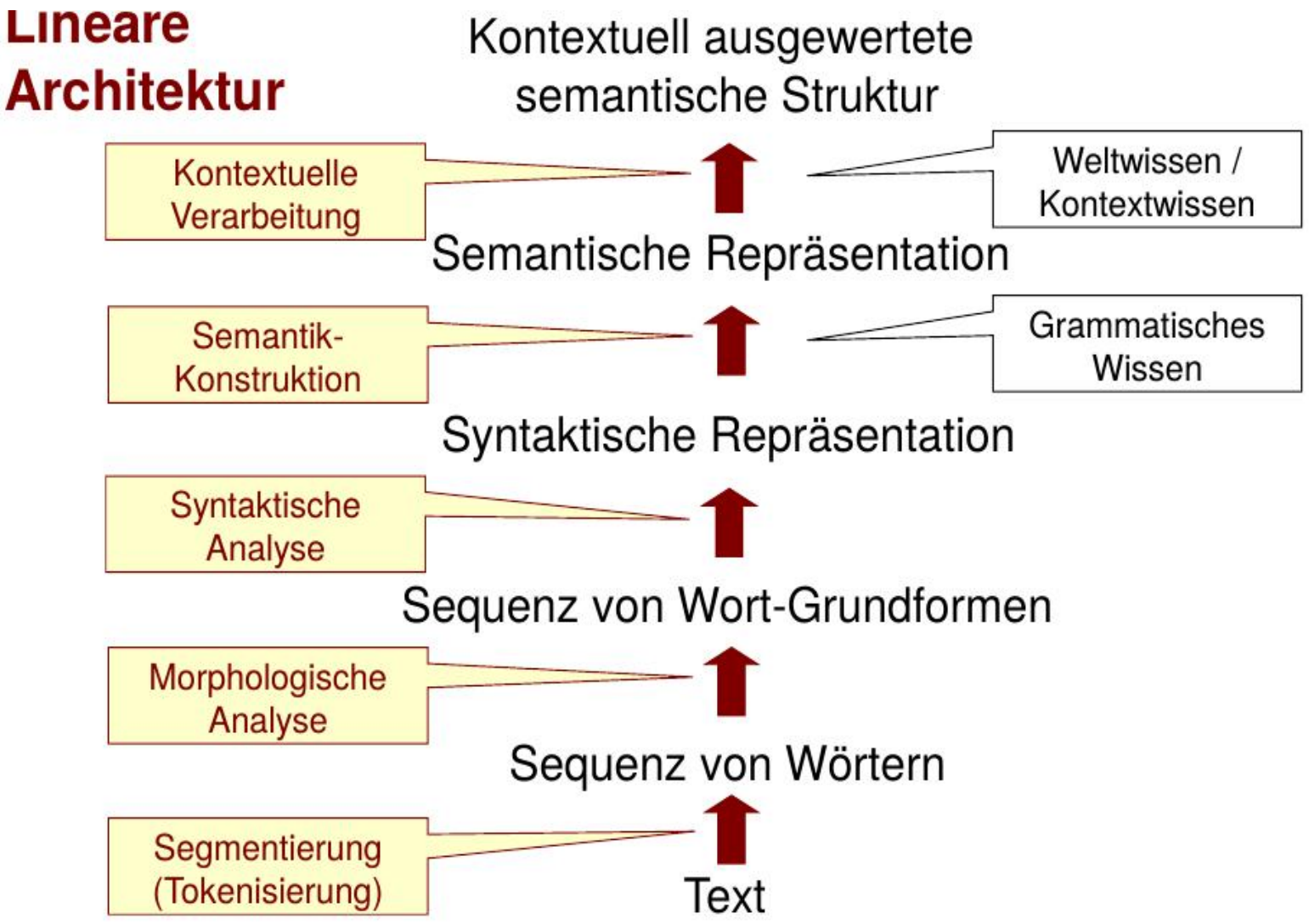
Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.

- der Relativsatz könnte jede der vier Nominalphrasen im Plural modifizieren (4);
- *die* als auch *ihre Männer* kann Subjekt des Relativsatzes sein (2);
- das Possessivpronomen *ihre* kann auf jede der Nominalphrasen referieren (4);
- *Montagen* hat eine zweite Lesart als Nominalisierung von montieren (2);
- *Hauptinsel* kann im Genitiv zu der vorangegangenen NP gehören oder im Dativ die Käuferin bezeichnen (2);
- die drei Präpositionalphrasen des Relativsatzes knnen sich in insgesamt sieben Kombinationen mit den jeweils vorhergehenden NPs oder mit dem Verb verbinden (7);
- *Verkauften* zeigt wieder die Ambiguität zwischen Präteritum und Konjunktiv auf (2).

Quelle: Hans Uszkoreit

Übersicht

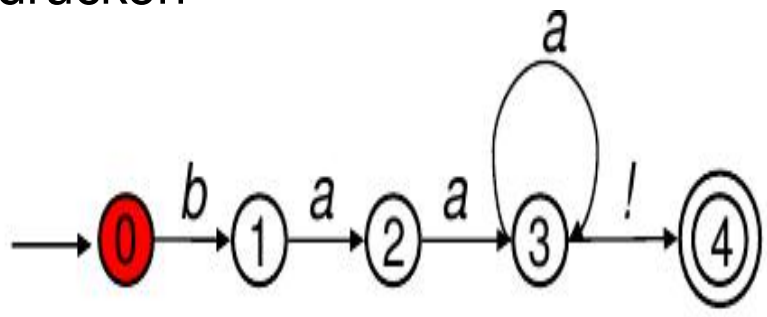
Lineare Architektur



Teil 0: Kognitive Grundlagen (nächstes Mal)

- Was macht Sprache als Kommunikationssystem besonders?
- Vergleich mit tierischer Kommunikation
- Beste Behandlung von sprachlichen Phänomenen: Mustererkennung oder regelbasiert?

- Text (und einzelne Worte) als Sequenz von Zeichen
- **Reguläre Ausdrücke:** Muster für Text, nach denen man (mit jeder Programmiersprache) suchen kann
- Beispiel: Baa+! $\longrightarrow$ Baa!, Baaa! ...
- Endliche Automaten als formaler Mechanismus hinter regulären Ausdrücken



- **Tokenisierung:** wo fangen Wörter und Sätze an, wo hören sie auf?
- Effiziente Algorithmen für Stringvergleich; dynamic programming

Teil II: Sequenzmodellierung, Wortverteilungen

- Wie sehen Wortverteilungen in Texten aus?
- Plausibilität von Wort- oder Buchstabenfolgen: *drink white . . .*
- **Smoothing**: Was mache ich mit ungesehenen Wortfolgen?
- Anwendung: Language Identification
- Hintergrund: Wiederholung Wahrscheinlichkeitstheorie sowie Einführung in Informationstheorie)

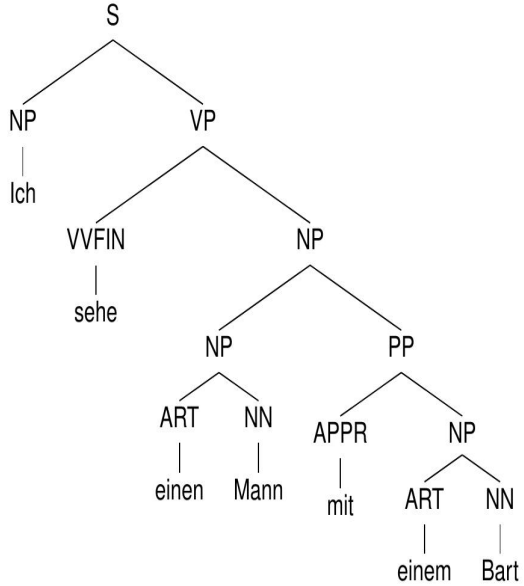
Teil III: Textklassifikation

- Was kann man schon mit nur Worten und ein paar Wortfolgen tun? **Textklassifikation!**
- Klassifiziere Texte nach Inhalt, Genre, Autor, Meinung
- Hintergrund: maschinelles Lernen

Just came out of the theater and I'm literally blowing away! As a moviegoer and movie lover looking for a good entertaining is simply irresistible not to like this movie even just a little.

i love james bond,I've seen all the films, and i can say this is the worst one, dull , meandering script, at times i had no idea what the plot was. lots of confusiion

- In welche Klassen (z.B. Nomen, Verben, Adjektive) kann ich Wörter einteilen? Wie kann ich diese Klassen bei Worten in einem Text automatisch zuordnen?
- Wie kreiiere ich Syntaxbäume, automatisch und manuell?



- Wie behandle ich syntaktische Ambiguitäten?
- Methoden: formale Sprachen, Grammatiken, Suchalgorithmen, Hidden-Markov-Modelle, probabilistische Grammatiken, dynamische Programmierung

- **Distributionelle lexikalische Semantik:** Kann ich die Vorkommen eines Wortes in einem Textkorpus nutzen, um etwas über dessen Bedeutung zu erfahren? Um Ähnlichkeiten zwischen Worten zu berechnen?
- Wahrscheinlich nicht: **Satzsemantik**: Wie komme ich von der Wortsemantik zur Satzsemantik? Semantische Rollen.
- Hintergrund: lineare Algebra, sparse and dense word embeddings, clustering

Teil VI: Diskurs sowie Anwendungsüberblick

- **Koreferenzresolution:** *Peter's car is in the garage. It is red.*
- Hintergrund: linguistische Constraints, Suchalgorithmen, Graphen
- Überblick über einige Anwendungen (z.B. IR) als Vorschau auf weitere Semester

Aufgaben

- Mindestens eine der Anwendungen oder Demos ausprobieren
- Kapitel 1 in Jurafsky und Martin (2nd edition) lesen