NLP and the Web – WS 2024/2025

Menti.com
2848 6424

Lecture 1
Introduction

Dr. Thomas Arnold
Hovhannes Tamoyan
Kexin Wang

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Introduction: Teaching Staff

Menti.com
2848 6424

Dr. Thomas Arnold
Lectures

Hovhannes Tamoyan
Practice Class

Kexin Wang
Practice Class

Outline

Menti.com
2848 6424

UKP Lab: profile and projects

Administrative course issues

NLP 4 Web Introduction

NLP Basics / Linguistic Analysis

Who Are We?

Menti.com
2848 6424

- 1 Professor, ~5 Postdocs, ~35 Doctoral Researchers
- We mainly work in natural language processing (NLP)
- Research areas (growing every day!)

Deep Learning for NLP

Knowledge Graphs

Argument Mining

Interactive AI and NLP

Content Analytics for the Social Good

Writing Assistance and Language Learning

Teaching Concept – UKP (Lectures)

Menti.com
2848 6424

	Winter Term	Summer Term
Introductory		Information Management
Application Oriented	NLP and the Web	Ethics in NLP
Advanced		Deep Learning for NLP

Teaching Concept – UKP (Seminars & Projects)

Menti.com
2848 6424

Software Project

(irregular schedule)

Data Analysis Software Project for Natural Language

Winter 2023/24: Various Projects
Winter 2024/25: Various Projects

Regular Seminar

Text Analytics / Large Language Models

Winter 2023/24: Generative AI
Summer 2024: LLMs for Mental Health
Winter 2024/25: Understanding LLMs

Complementary Lectures and Seminars

Menti.com
2848 6424

- **Machine Learning**
  - Einführung in die künstliche Intelligenz (Kersting)
  - Data Mining und maschinelles Lernen (Kersting)
  - Deep Learning (Kersting)

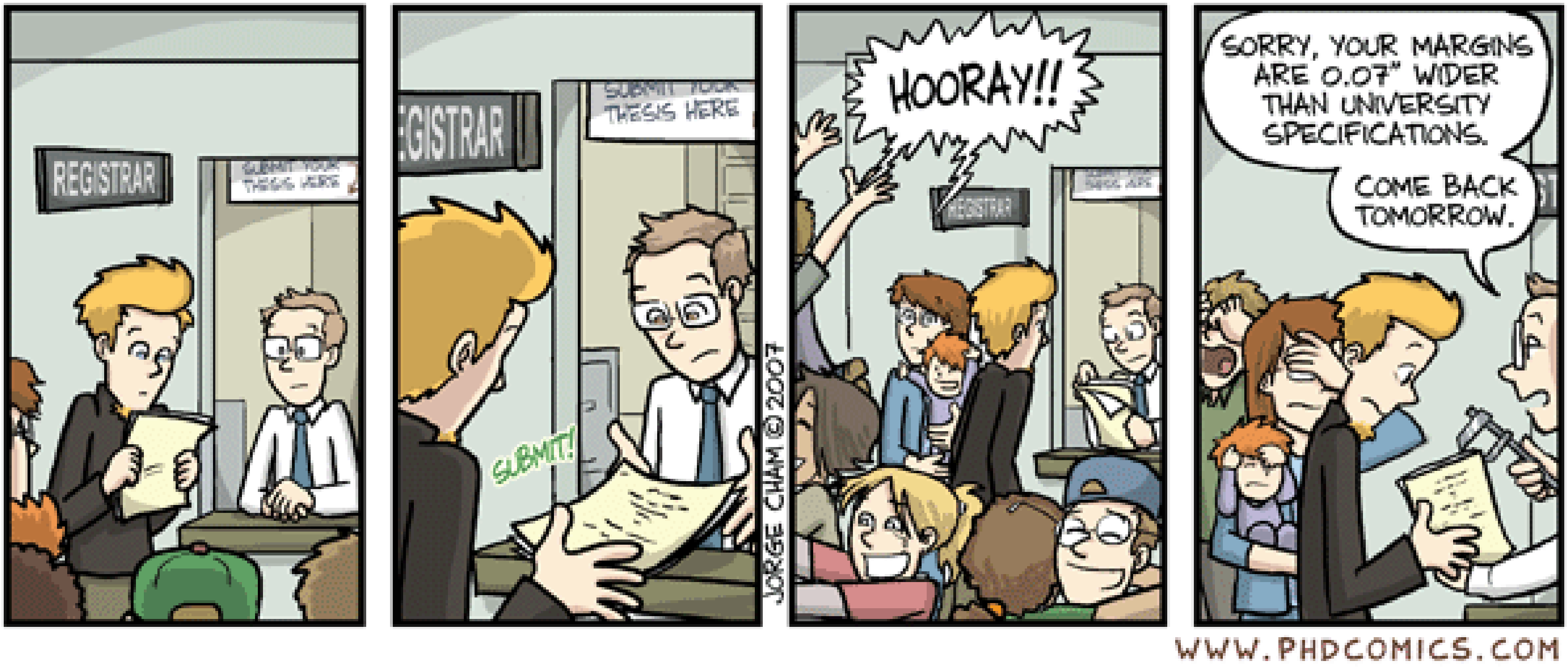
- **Computer Vision**
  - Computer Vision 1 and 2 (Roth)

- **Natural Language Processing**
  - Deep Learning for NLP
  - Ethics in NLP

Teaching Concept – UKP (PhD)

Menti.com
2848 6424

- Get involved early (HiWi, B.Sc. thesis, M.Sc. thesis)



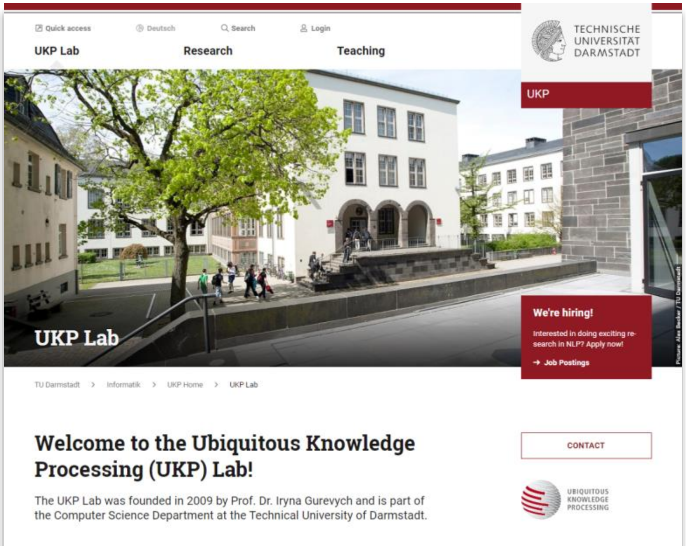
More information

Menti.com
8411 3758

UBIQUITOUS
KNOWLEDGE
PROCESSING

- Website:
  www.ukp.tu-darmstadt.de
- GitHub:
  www.github.com/UKPLab
- Social Media:

@UKPLab


Quick access
Deutsch
Search
Login
UKP Lab
Research
Teaching
TECHNISCHE UNIVERSITÄT DARMSTADT
UKP
UKP Lab
Job Postings
Picture: Alex Becker / TU Darmstadt
TU Darmstadt > Informatik > UKP Home > UKP Lab
CONTACT
UBIQUITOUS KNOWLEDGE PROCESSING

We're hiring!

Interested in doing exciting research in NLP? Apply now!

Welcome to the Ubiquitous Knowledge Processing (UKP) Lab!

The UKP Lab was founded in 2009 by Prof. Dr. Iryna Gurevych and is part of the Computer Science Department at the Technical University of Darmstadt.

Outline

Menti.com
8411 3758

UKP Lab: profile and projects

Administrative course issues

NLP 4 Web Introduction

NLP Basics / Linguistic Analysis

Course Goals

Menti.com
2848 6424

- Learn the basic principles underlying **NLP systems**
- Two big NLP topics:
    - Information Retrieval (IR)
    - Large Language Model (LLM) Applications
- Gain insight into **open research problems** in natural language processing

Why Care?

Menti.com
2848 6424

Information Overload

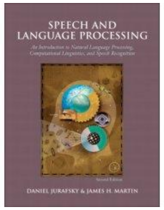
Business Intelligence

Need for Robust, Intelligent Systems

Textbook

Menti.com
2848 6424

Constantly updated:



- Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Daniel Jurafsky and James H. Martin. 3nd edition, 2023 (draft).
  - https://web.stanford.edu/~jurafsky/slp3/

General Information

Menti.com
2848 6424

- All lectures and practice classes will be in person
  Lectures: Tuesdays 13:30 – 15:10, S306 / 051
  Practice Class: Thursdays 16:15 – 17:55, S103 / 221

- All slides, handouts, readings etc. can be found on the Moodle e-Learning platform

- We also use Moodle as a central point for announcements and questions

- **Please use the Moodle forum!**

General Information – Practice Class

Menti.com
2848 6424

- In the **practice classes**, you will work on programming exercises
  - Programming language is Python
  - First practice session will include a brief introduction to Python
  - This will give you some practical experience in NLP
  - Practice class topics are **relevant for the exam**! (including Python)

- In addition, there are **homework assignments** for an exam bonus:
  - Assignments will be bi-weekly – 6 exercises in total
  - Each assignment is worth a maximum of 20 points
  - If you get >= 75% of the points (>= 90 points), you get a bonus
    - You can improve your grade by 0.3/0.4 IFF you pass the exam without bonus

General Information – Practice Class

Menti.com
2848 6424

- First class: **October 24th (no practice class this week)**

- Details will be announced in moodle
- If you need additional help regarding the practice class, use the Moodle forum

The assignments will require a significant amount of time, so start earlier than the day before submission.

Final exam

Menti.com
2848 6424

Tuesday, 25.02.2025, 15:00
More info be announced in Moodle

- **Allowed:** Non-programmable calculator, no other material
- **Content:** lecture, readings, practice class

Syllabus (tentative)

Menti.com
2848 6424

Nr.	Lecture
01	**Introduction / NLP basics**
02	Foundations of Text Classification
03	IR – Introduction, Evaluation
04	IR – Word Representation, Data Collection
05	IR – Re-Ranking Methods
06	IR – Language Domain Shifts, Dense / Sparse Retrieval
07	LLM – Language Modeling Foundations
08	LLM – Neural LLM, Tokenization
09	LLM – Transformers, Self-Attention
10	LLM – Adaption, LoRa, Prompting
11	LLM – Alignment, Instruction Tuning
12	LLM – Long Contexts, RAG
13	LLM – Scaling, Computation Cost
14	Review & Preparation for the Exam

Warm up

Menti.com
2848 6424

Now it is your turn:

Which degree programme are you studying?

- Computer Science?
- Bachelor?
- Master?
- Other disciplines?

Warm up

Menti.com
2848 6424

Now it is your turn:

Which other UKP courses did you already attend?

- FoLT
- Ethics in Natural Language Processing
- Deep Learning for NLP
- Data Analysis Software Project
- Text Analytics / LLM Seminar

Outline

Menti.com
2848 6424

UKP Lab: profile and projects

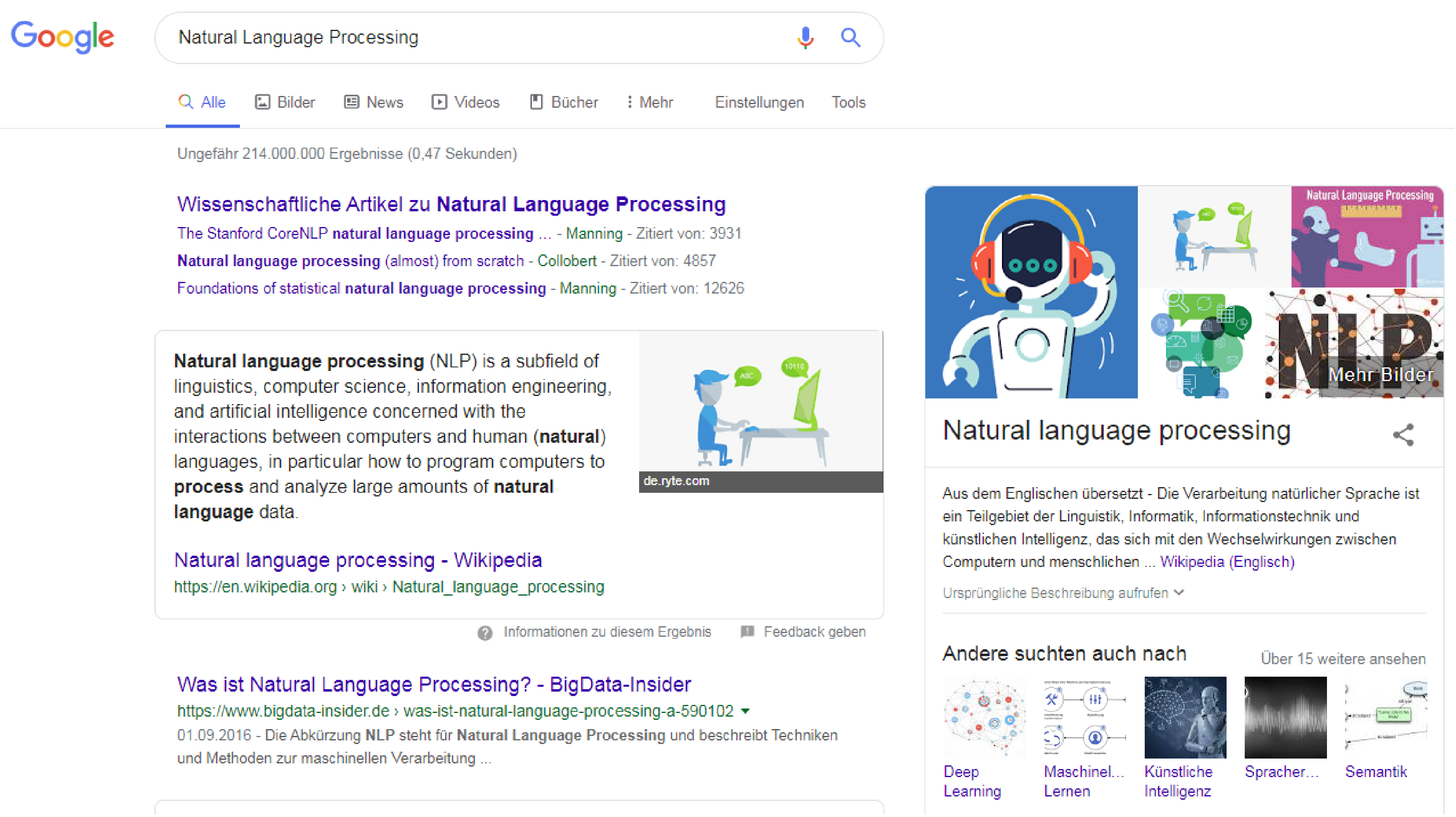
Administrative course issues

NLP 4 Web Introduction

NLP Basics / Linguistic Analysis

NLP in the Web – Search Engines

Menti.com
2848 6424



Ungefähr 214.000.000 Ergebnisse (0,47 Sekunden)

Wissenschaftliche Artikel zu **Natural Language Processing**

The Stanford CoreNLP **natural language processing** ... - Manning - Zitiert von: 3931

Natural language processing (almost) from scratch - Collobert - Zitiert von: 4857

Foundations of statistical **natural language processing** - Manning - Zitiert von: 12626

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (**natural**) languages, in particular how to program computers to **process** and analyze large amounts of **natural language** data.

Natural language processing - Wikipedia

https://en.wikipedia.org › wiki › Natural_language_processing

Informationen zu diesem Ergebnis   Feedback geben

Was ist Natural Language Processing? - BigData-Insider

https://www.bigdata-insider.de › was-ist-natural-language-processing-a-590102 ▾

01.09.2016 - Die Abkürzung **NLP** steht für **Natural Language Processing** und beschreibt Techniken und Methoden zur maschinellen Verarbeitung ...

Natural language processing

Aus dem Englischen übersetzt - Die Verarbeitung natürlicher Sprache ist ein Teilgebiet der Linguistik, Informatik, Informationstechnik und künstlichen Intelligenz, das sich mit den Wechselwirkungen zwischen Computern und menschlichen ... Wikipedia (Englisch)

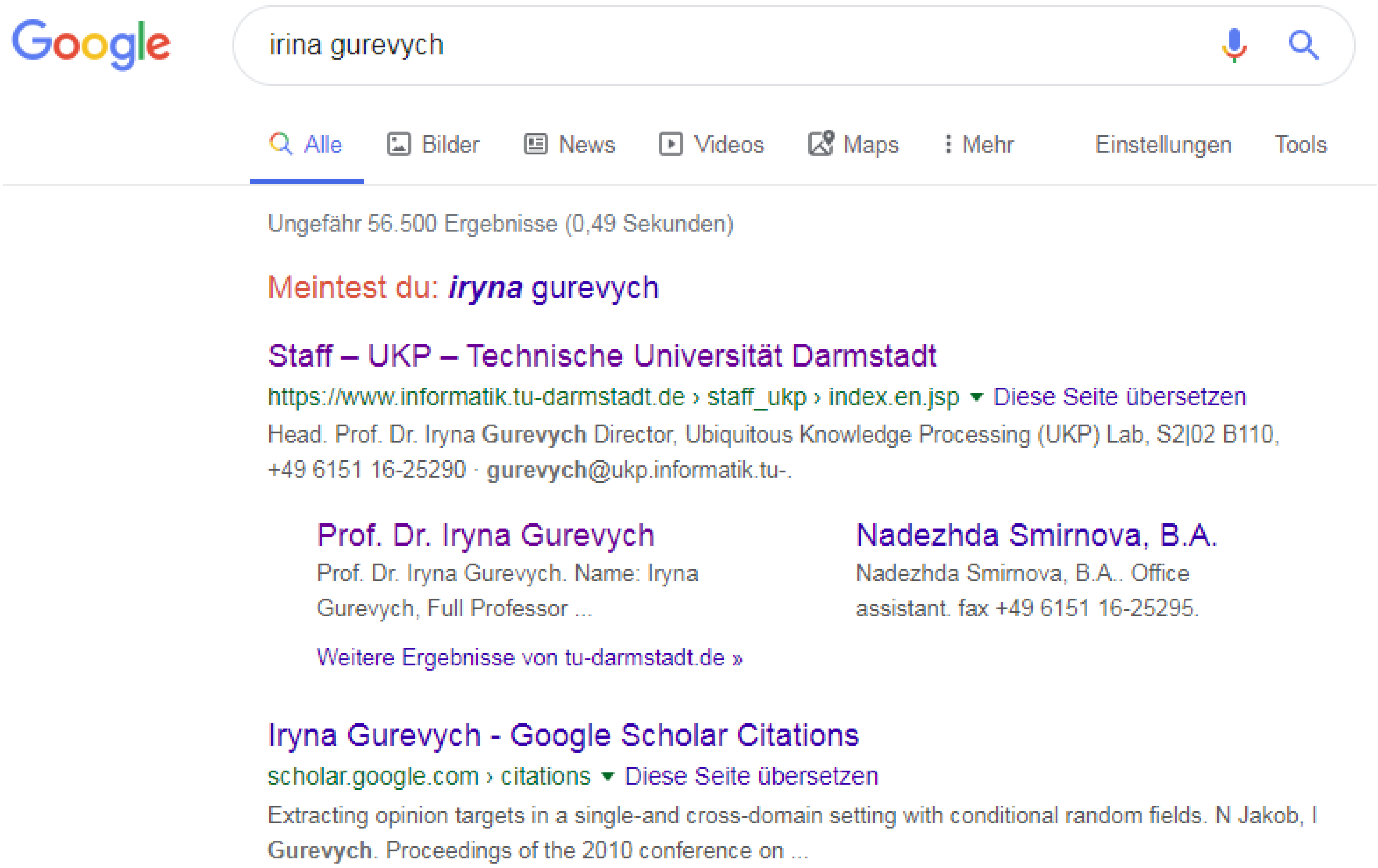
Ursprüngliche Beschreibung aufrufen ⌄

Andere suchten auch nach   Über 15 weitere ansehen

Deep Learning
Maschinel... Lernen
Künstliche Intelligenz
Spracher...
Semantik

NLP in the Web – Spelling Correction

Menti.com
2848 6424



irina gurevych

Ungefähr 56.500 Ergebnisse (0,49 Sekunden)

Meintest du: ***iryna*** gurevych

Staff – UKP – Technische Universität Darmstadt
https://www.informatik.tu-darmstadt.de › staff_ukp › index.en.jsp ▾ Diese Seite übersetzen
Head. Prof. Dr. Iryna **Gurevych** Director, Ubiquitous Knowledge Processing (UKP) Lab, S2|02 B110, +49 6151 16-25290 · **gurevych**@ukp.informatik.tu-.

Prof. Dr. Iryna Gurevych
Prof. Dr. Iryna Gurevych. Name: Iryna Gurevych, Full Professor ...

Nadezhda Smirnova, B.A.
Nadezhda Smirnova, B.A.. Office assistant. fax +49 6151 16-25295.

Weitere Ergebnisse von tu-darmstadt.de »

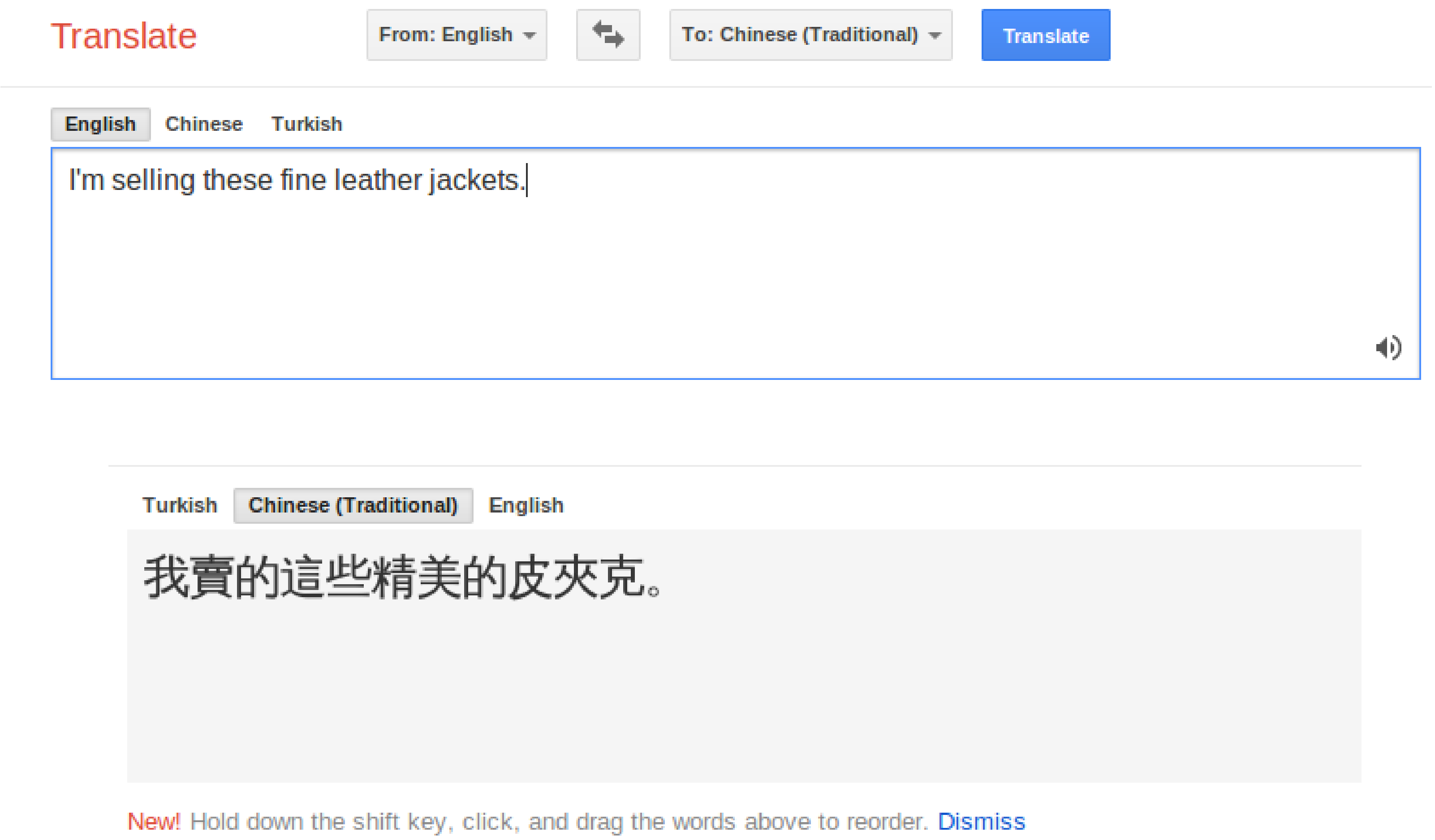
Iryna Gurevych - Google Scholar Citations
scholar.google.com › citations ▾ Diese Seite übersetzen
Extracting opinion targets in a single-and cross-domain setting with conditional random fields. N Jakob, I **Gurevych**. Proceedings of the 2010 conference on ...

Question Answering

Menti.com
2848 6424

NLP in the Web – Machine Translation

Menti.com
2848 6424



I'm selling these fine leather jackets.

我賣的這些精美的皮夾克。

New! Hold down the shift key, click, and drag the words above to reorder. Dismiss

NLP in the Web – Speech Recognition

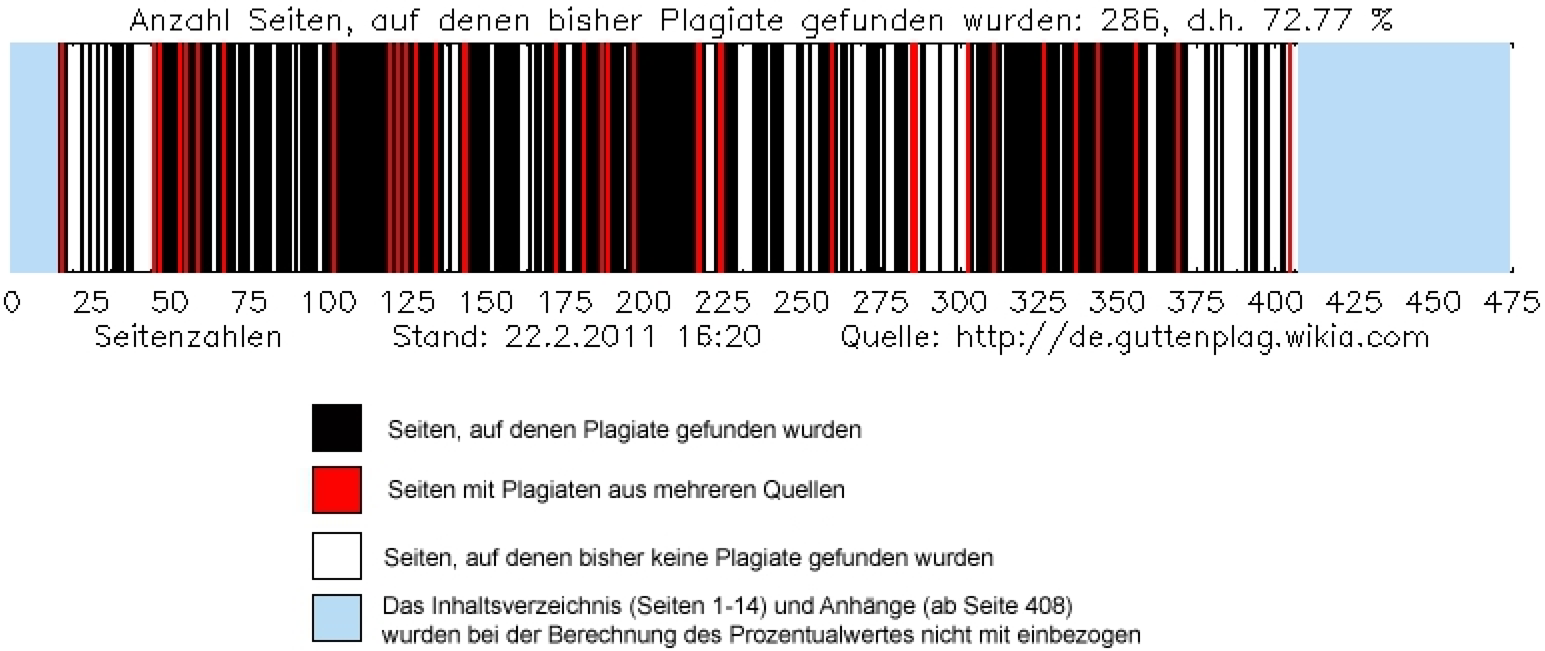
Menti.com
2848 6424

« Please, Siri, open my contacts »
I'm afraid I can't do that, Dave.

NLP in the Web – Plagiarism Detection

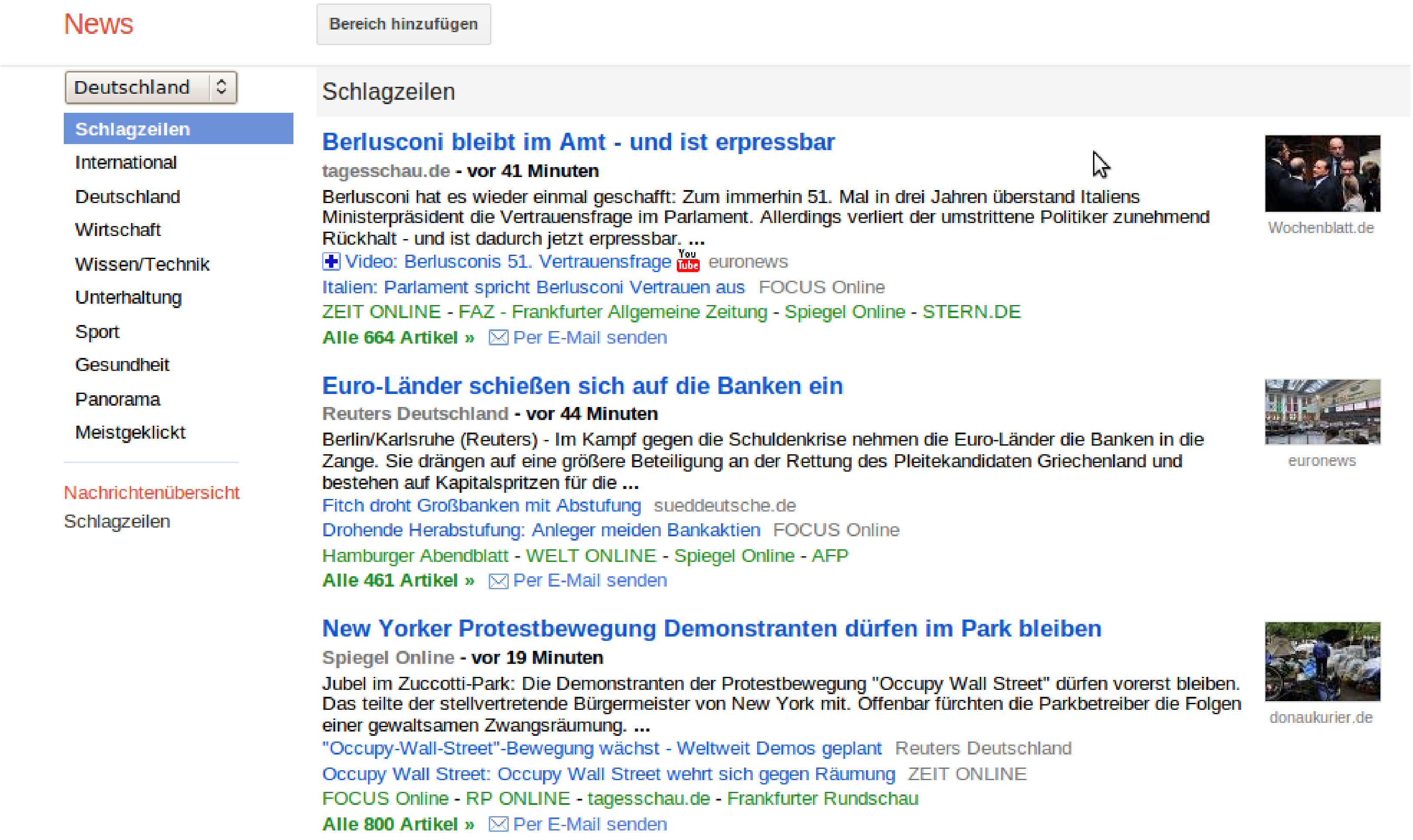
Menti.com
2848 6424

http://de.guttenplag.wikia.com/



NLP in the Web – Summarization

Menti.com
2848 6424



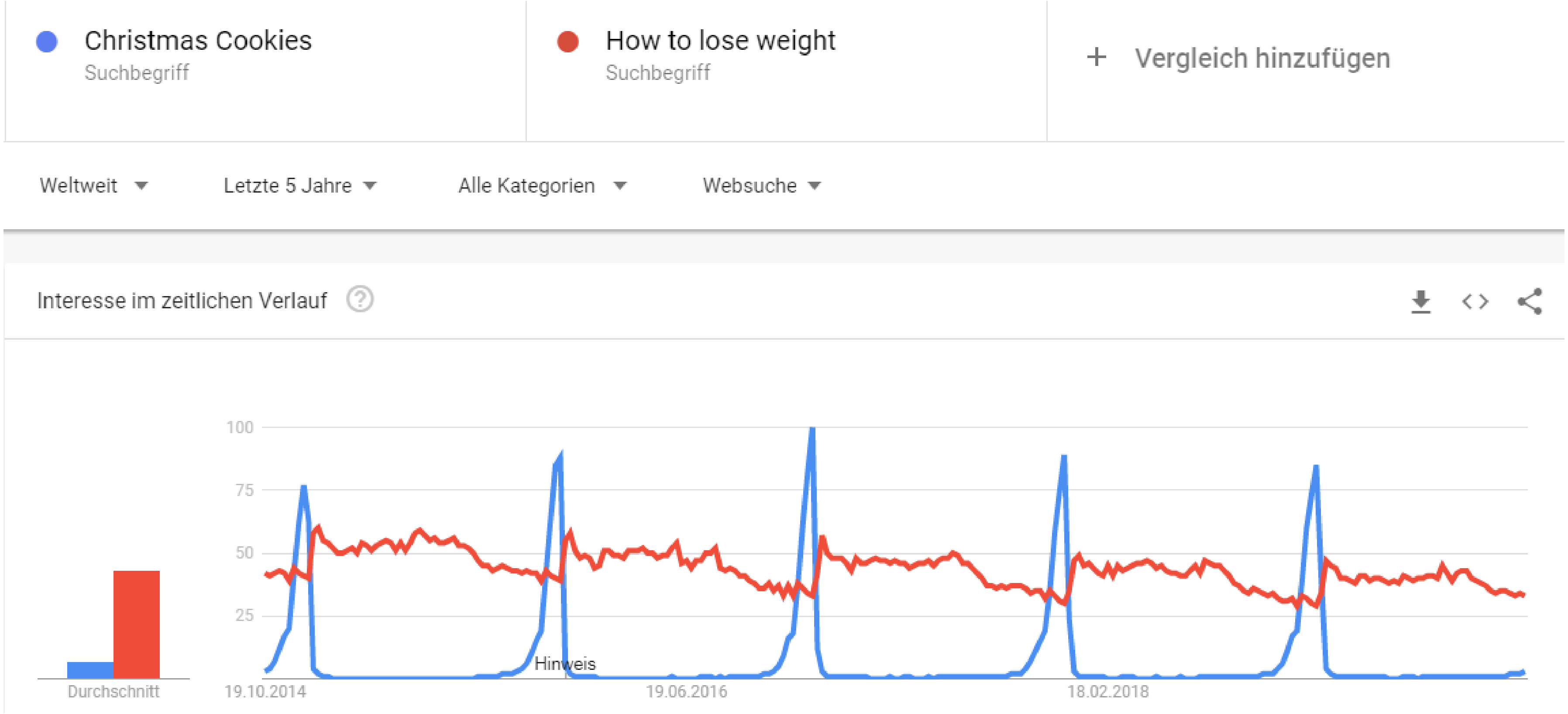
Berlusconi bleibt im Amt - und ist erpressbar
tagesschau.de - **vor 41 Minuten**
Berlusconi hat es wieder einmal geschafft: Zum immerhin 51. Mal in drei Jahren überstand Italiens Ministerpräsident die Vertrauensfrage im Parlament. Allerdings verliert der umstrittene Politiker zunehmend Rückhalt - und ist dadurch jetzt erpressbar. **...**
Video: Berlusconis 51. Vertrauensfrage euronews
Italien: Parlament spricht Berlusconi Vertrauen aus FOCUS Online
ZEIT ONLINE - FAZ - Frankfurter Allgemeine Zeitung - Spiegel Online - STERN.DE
Alle 664 Artikel » ✉ Per E-Mail senden

Euro-Länder schießen sich auf die Banken ein
Reuters Deutschland - **vor 44 Minuten**
Berlin/Karlsruhe (Reuters) - Im Kampf gegen die Schuldenkrise nehmen die Euro-Länder die Banken in die Zange. Sie drängen auf eine größere Beteiligung an der Rettung des Pleitekandidaten Griechenland und bestehen auf Kapitalspritzen für die **...**
Fitch droht Großbanken mit Abstufung sueddeutsche.de
Drohende Herabstufung: Anleger meiden Bankaktien FOCUS Online
Hamburger Abendblatt - WELT ONLINE - Spiegel Online - AFP
Alle 461 Artikel » ✉ Per E-Mail senden

New Yorker Protestbewegung Demonstranten dürfen im Park bleiben
Spiegel Online - **vor 19 Minuten**
Jubel im Zuccotti-Park: Die Demonstranten der Protestbewegung "Occupy Wall Street" dürfen vorerst bleiben. Das teilte der stellvertretende Bürgermeister von New York mit. Offenbar fürchten die Parkbetreiber die Folgen einer gewaltsamen Zwangsräumung. **...**
"Occupy-Wall-Street"-Bewegung wächst - Weltweit Demos geplant Reuters Deutschland
Occupy Wall Street: Occupy Wall Street wehrt sich gegen Räumung ZEIT ONLINE
FOCUS Online - RP ONLINE - tagesschau.de - Frankfurter Rundschau
Alle 800 Artikel » ✉ Per E-Mail senden

NLP in the Web – Diachronic Analysis

Menti.com
2848 6424



● Christmas Cookies
Suchbegriff

● How to lose weight
Suchbegriff

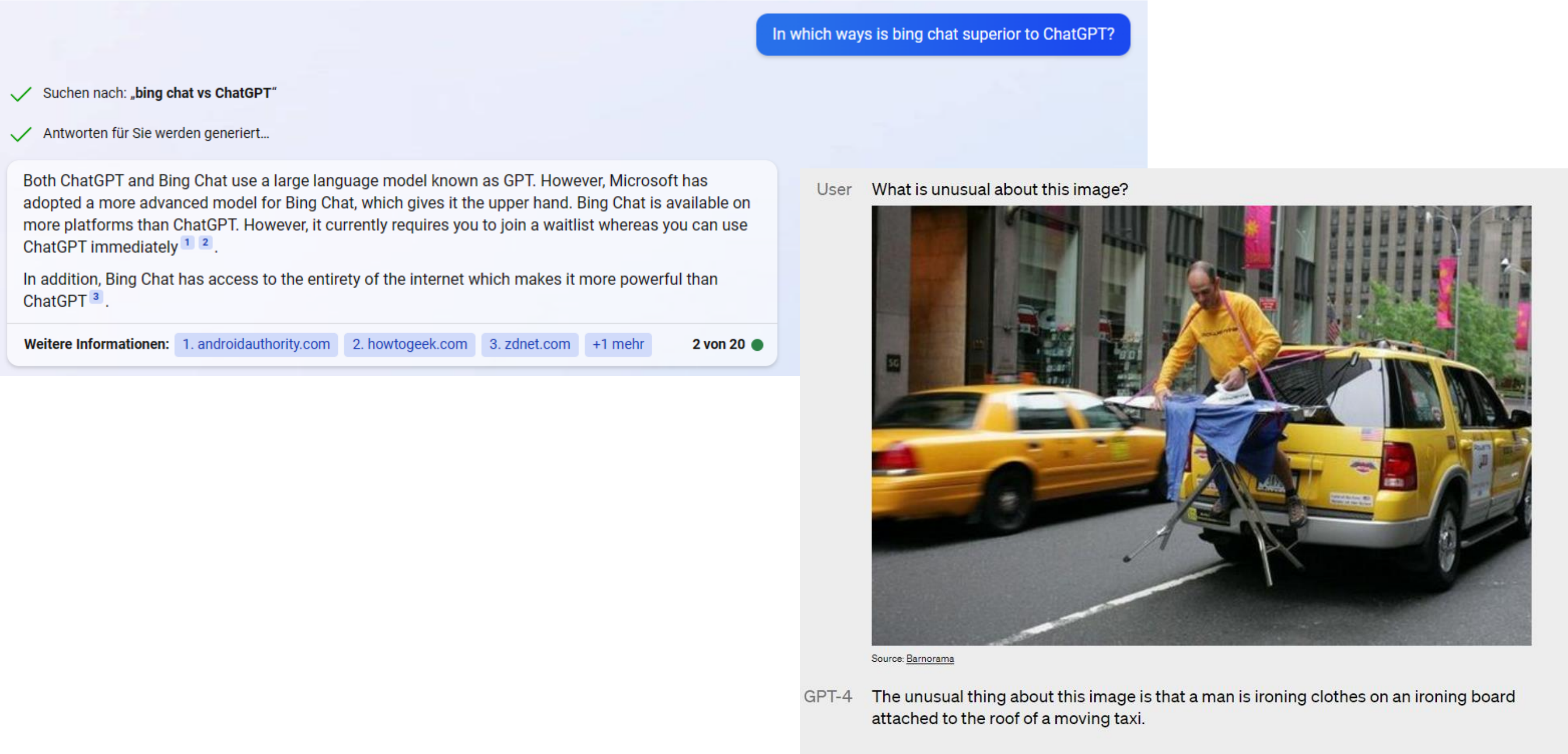
+ Vergleich hinzufügen

Weltweit ▾ Letzte 5 Jahre ▾ Alle Kategorien ▾ Websuche ▾

Interesse im zeitlichen Verlauf

NLP in the Web – Text Generators

Menti.com
2848 6424



In which ways is bing chat superior to ChatGPT?

✓ Suchen nach: „**bing chat vs ChatGPT**"

✓ Antworten für Sie werden generiert…

Both ChatGPT and Bing Chat use a large language model known as GPT. However, Microsoft has adopted a more advanced model for Bing Chat, which gives it the upper hand. Bing Chat is available on more platforms than ChatGPT. However, it currently requires you to join a waitlist whereas you can use ChatGPT immediately [1] [2].

In addition, Bing Chat has access to the entirety of the internet which makes it more powerful than ChatGPT [3].

Weitere Informationen: 1. androidauthority.com 2. howtogeek.com 3. zdnet.com +1 mehr **2 von 20** ●

User What is unusual about this image?

Source: Barnorama

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Natural Language Processing and the Web

Menti.com
2848 6424

- The web is an **application area** for NLP, e.g.:
  - Information retrieval:
    - Search engines
    - Question answering
    - News aggregation
  - Recommender Systems
  - Chatbots…
- Web is a **resource** to improve the quality of NLP, e.g.:
  - Web as a corpus
  - Analyzing web-based knowledge repositories
    - Wikipedia
    - Wiktionary
  - Recognizing synonyms, paraphrases and the like

Challenges for NLP

Menti.com
2848 6424

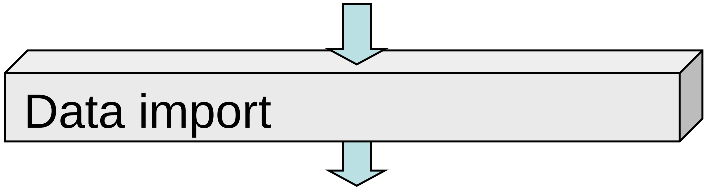
- How to remove noise, e.g. duplicates?
- How to assess the quality of content?
- How to integrate the content of heterogeneous and scattered nature?
- How to deal with errors, e.g. spelling or grammar errors?
- How to „clean“ the data?

Data Cleansing is Necessary

Menti.com
2848 6424

- User-generated content contains errors, smileys, abbreviations, etc.

```
Hi
        Micheal,
                           have u seen my
posting,last week u said that u
will look in to my problem thsi week.can i ask u
now?
```



Data import

Data cleansing

Outline

Menti.com
2848 6424

UKP Lab: profile and projects

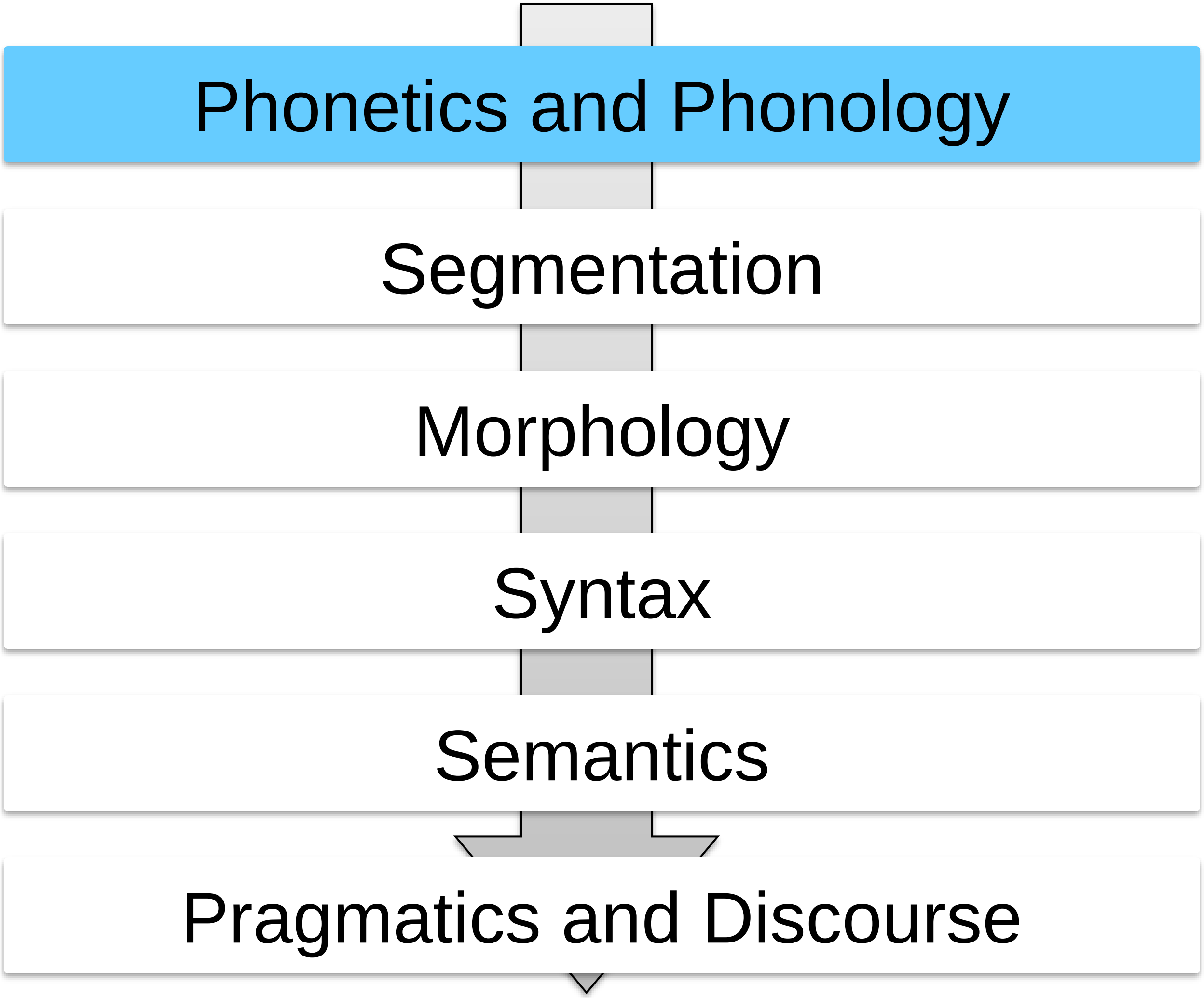
Administrative course issues

NLP 4 Web Introduction

NLP Basics / Linguistic Analysis

Analysis Levels in Language Understanding

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

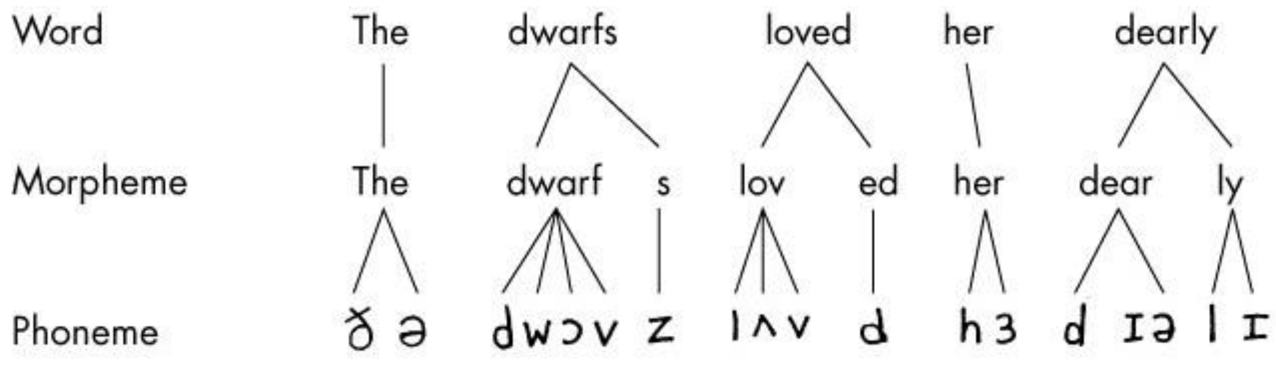
Syntax

Semantics

Pragmatics and Discourse

Phonetics and Phonology

Menti.com
2848 6424



(c) David Groome, 2006

Homophones

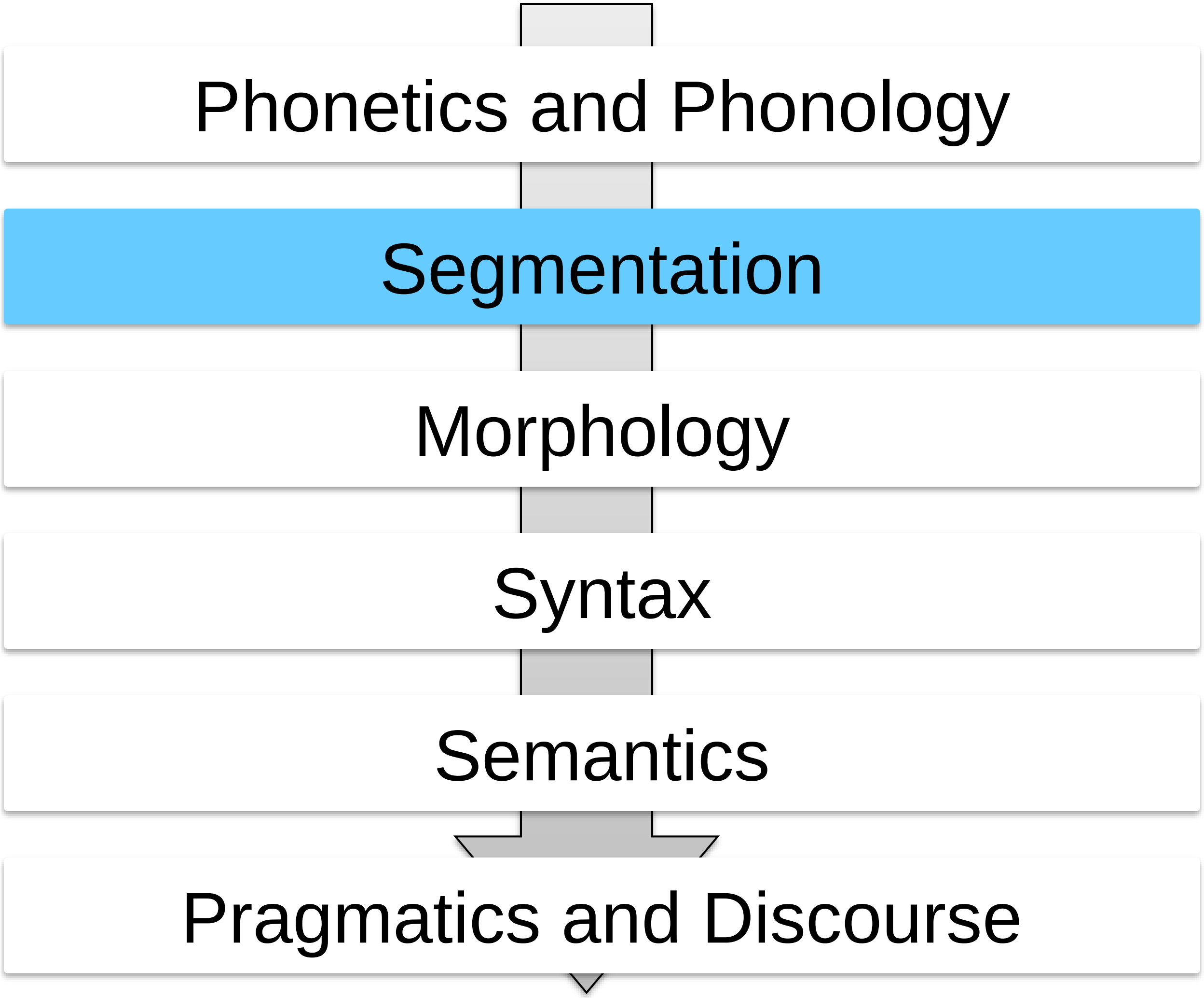
/naɪt/

night

knight

Analysis Levels in Language Understanding

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

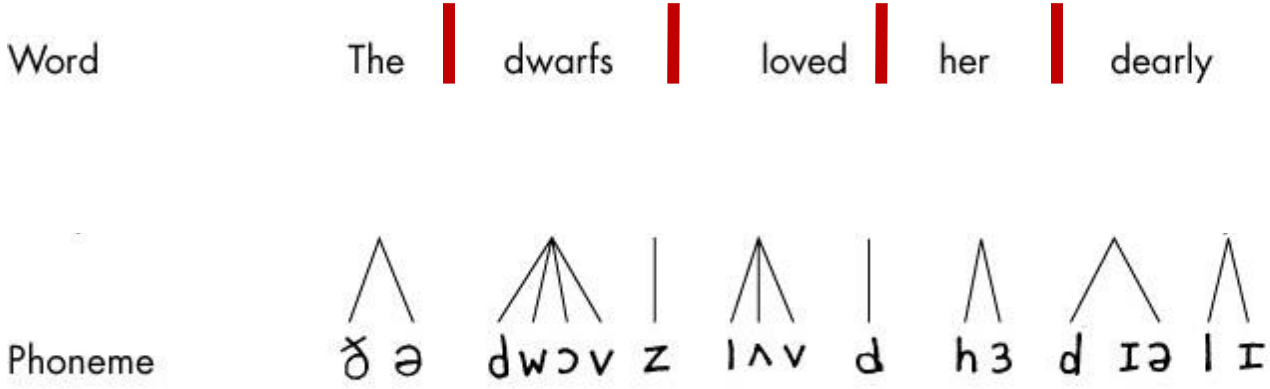
Syntax

Semantics

Pragmatics and Discourse

Segmentation

Menti.com
2848 6424



(c) David Groome, 2006

Tokenization

Menti.com
2848 6424

- Segmenting an input stream into an ordered sequence of units is called **tokenization**.
- A token can correspond to an inflected word form or sub-word units, and may be subject to a subsequent morphological analysis.
- Tokens include punctuation!

- A system which splits texts into tokens is called a **tokenizer**

A very simple example:

- Input text:
  John likes Mary and Mary likes John.
- Tokens:
  {"John", "likes", "Mary", "and", "Mary", "likes", "John", "."}

Tokenization

Menti.com
2848 6424

English Example

- Mr. Sherwood said, reaction to Sea Containers‘ proposal has been „very positive.“ In New York Stock Exchange composite trading yesterday, Sea Containers closed at $62.625, up 62.5 cents.

Where could be problems for a tokenizer?

Tokenization

Menti.com
2848 6424

English Example

- Mr. Sherwood said, reaction to Sea Containers‘ proposal has been „very positive.“ In New York Stock Exchange composite trading yesterday, Sea Containers closed at $62.625, up 62.5 cents.

- Split at whitespace characters?
  `cents.` `said,` `positive.”` `$62.625,`

Tokenization Ambiguities

Menti.com
2848 6424

Period

- In most of the cases: Final sentence punctuation symbol
- Part of an abbreviation, e.g. F.D.P.
- Numbers, ordinal numbers, e.g.: 21., numbers with fractions, e.g. 1.543
- References to resources locators, e.g.: www.apple.com
- To complicate things, if a sentence ends with an abbreviation which ends with a period, only one period is written. “I go to Apple, Inc.”
- …

Whitespace character

- Part of numbers, e.g. “1 543”
- No segmentation character in multi-word expressions
  - “New York”

Ambiguities

Menti.com
2848 6424

Comma

- Part of numbers, e.g. 1,543

Single quote

- Within tokens to mark contractions and elisions, e.g. English: *don’t, won’t, you’ve, James’ new hat*; German: *Ich hab’s!*
- Part of a token in French, e.g. *aujourd´hui*
- But in **most cases:** Enclosing quoted groups of words

Dash

- A delimiter, if it connects strings of digits, e.g. "see pages 100-101”
- In French: Signal a close connection between two tokens, e.g. verb and personal pronoun: *donne-le*
- In **most cases**, however, it is part of the token, e.g. *multi-word*

Tokenization in Other Languages

Menti.com
2848 6424

Chinese

爱国人

- No spaces
- Two possible segmentations, both of them are syntactically and semantically correct
- Disambiguation can only be done with contextual information

爱国/人

country-loving person

爱/国人

love country-person

Bird et al., NLP with Python, p.113

German Compounds

Menti.com
2848 6424

German

STAUBECKEN

- No spaces within noun compounds
- Two possible segmentations, both of them are syntactically and semantically correct
- Disambiguation can only be done with contextual information

STAU BECKEN

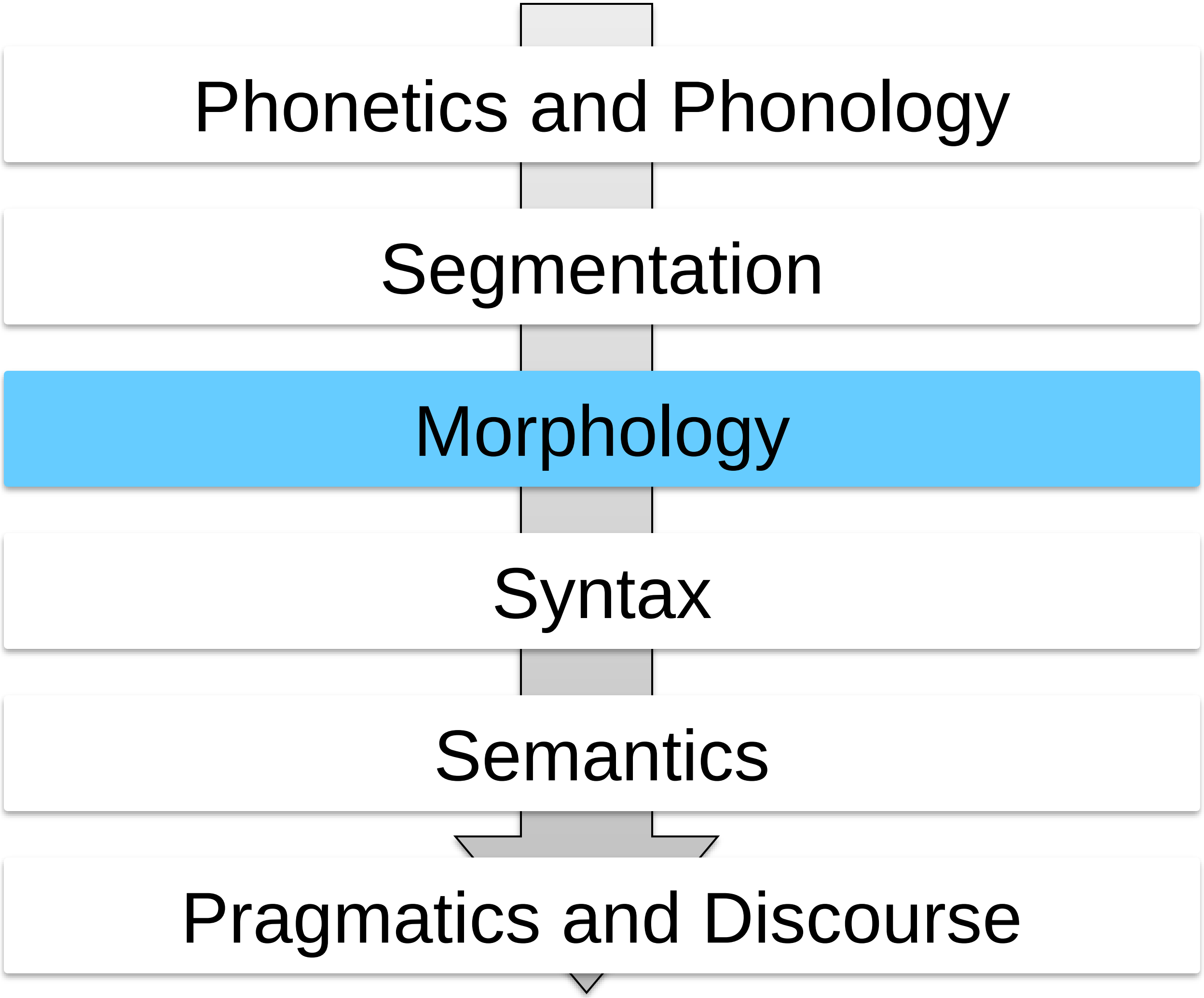
water reservoir

STAUB ECKEN

dusty corners

Analysis Levels in Language Understanding

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

Syntax

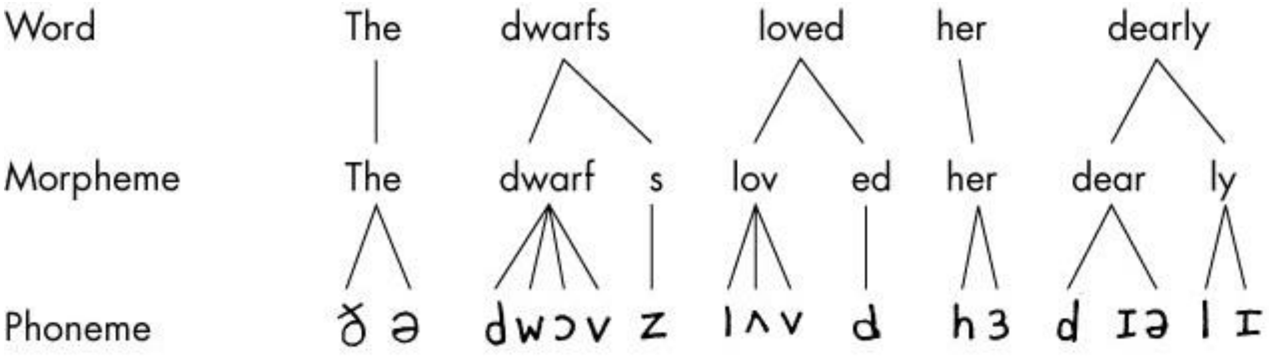
Semantics

Pragmatics and Discourse

Morphology

Menti.com
2848 6424

- Morphology is the branch of linguistics that studies word forms and word formation
- Words are composed of **morphemes**
- Morphemes are the **smallest meaning-bearing units**



(c) David Groome, 2006

Morphology

Menti.com
2848 6424

Words can be further decomposed into smaller units:

“pneumonoultramicroscopicsilicovolcanoconiosis”

lung disease caused by the inhalation of very fine silica dust found in volcanoes

Bases and Affixes

Menti.com
2848 6424

- Remember: Morphemes are the smallest meaning-bearing units
- Examples:
  - cats → cat (noun) + s (plural)
  - unknowingly → un + know + ing + ly
  - bedenken → be + denk + en
- Both *cat* and *cats* can be uttered in isolation but *s* cannot:
  -*s* is a **bound** morpheme

- Minimal **free** morphemes = **stems**
  - *cat* is a free morpheme
- Stems carry the main meaning of the word
- Affixes are bound morphemes

Types of Affixes

Menti.com
2848 6424

Suffixes: appear after the base

- cat + s, nice + ly

Prefixes: appear before the base

- un + true

Infixes: appear inside the base

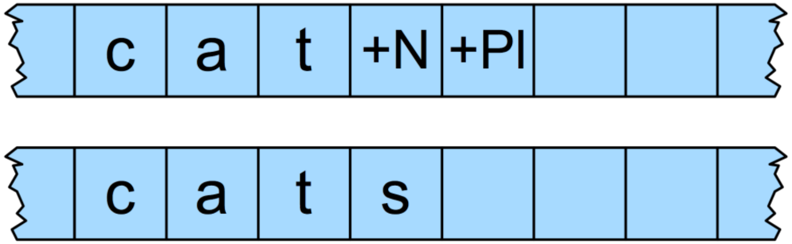
- fan + bloody + tastic

Circumfixes: appear on both sides of the base

- ge + sag + t

Morphological Normalization

Menti.com
2848 6424



- **Morphological normalization** consists in identifying a single canonical representative for morphologically related word-forms

Methods

- Stemming
- Lemmatization

Stemming

Menti.com
2848 6424

Stemming is an algorithmic approach to strip off the endings of words

sitting → sitt

anarchism, anarchy, anarchistic → anarchi

Objective: group words belonging to the same morphological family by transforming them into the same stemmed representation

- stemming does not distinguish between inflection and derivation
- the stems obtained do not necessarily correspond to a real word form

Well-known stemming algorithms for English have been developed by Lovins and Porter

Algorithmic Stemming Method

Menti.com
2848 6424

Stemming is rule-based. Example rules from Porter:

```
*ATIONAL -> *ATE (relational -> relate)

*[> 0 vowels] + ING -> * (monitoring -> monitor)

*SSES -> *SS (grasses -> grass)
```

Rule-based stemming methods are hard to create, often yield arbitrary distinctions, but can be executed very quickly at runtime.

Porter's Stemmer

Menti.com
2848 6424

Original Word	Stemmed Word
vision	vision
visible	visibl
visibility	visibl
visionary	visionari
visioner	vision
visual	visual

Stemming Errors

Menti.com
2848 6424

Under-stemming: remove too little

- adhere → adher
- adhesion → adhes

Over-stemming: remove too much

- appendicitis → append
- append → append

Problem with Stemming: Syntactic Ambiguity

Menti.com
2848 6424

Homographs: words which have the same spelling but different meanings

I **saw** the **saw**

Past form
of the verb
SEE

≠

Singular form
of the noun
SAW

Such cases **cannot** be properly dealt with by stemming only, the word's grammatical category also has to be identified

Lemmatization

Menti.com
2848 6424

- “undo” the inflectional changes of a base form
- Usually needs lexical resources and **part-of-speech tagging**
  - cats (NOUN) → cat
  - left (VERB) → leave
  - left (ADJ) → left

- Has to deal with **Irregularities**
  - sing, sang, sung → sing
  - indices → index
  - Bäume → Baum

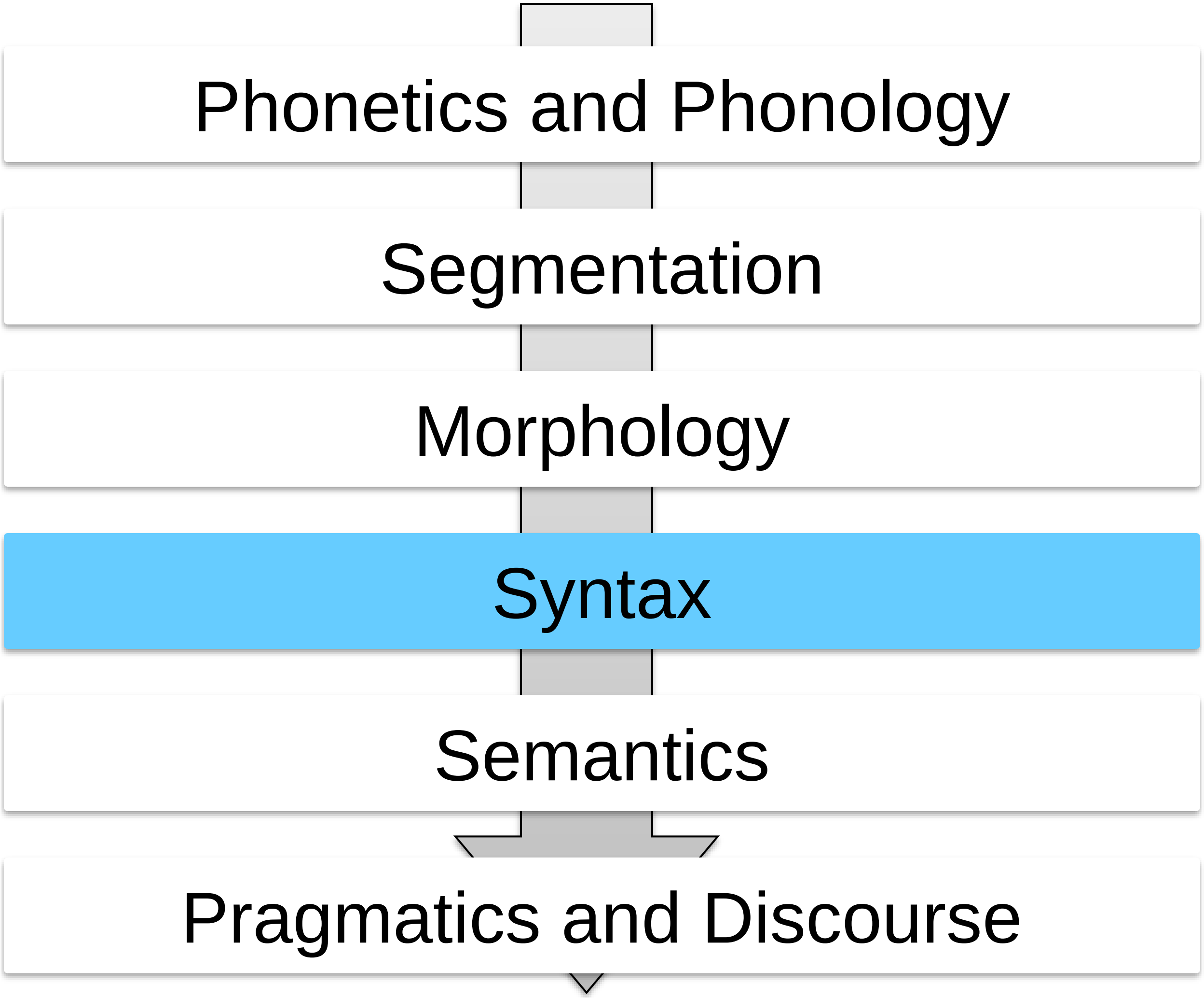
Stemming vs. Lemmatization

Menti.com
2848 6424

Original	Stemmed	Lemmatized
visibilities	visibl	visibility
adhere	adher	adhere
adhesion	adhes	adhesion
appendicitis	append	appendicitis
oxen	oxen	ox
indices	indic	index
swum	swum	swim

Analysis Levels in Language Understanding

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

Syntax

Semantics

Pragmatics and Discourse

Syntax

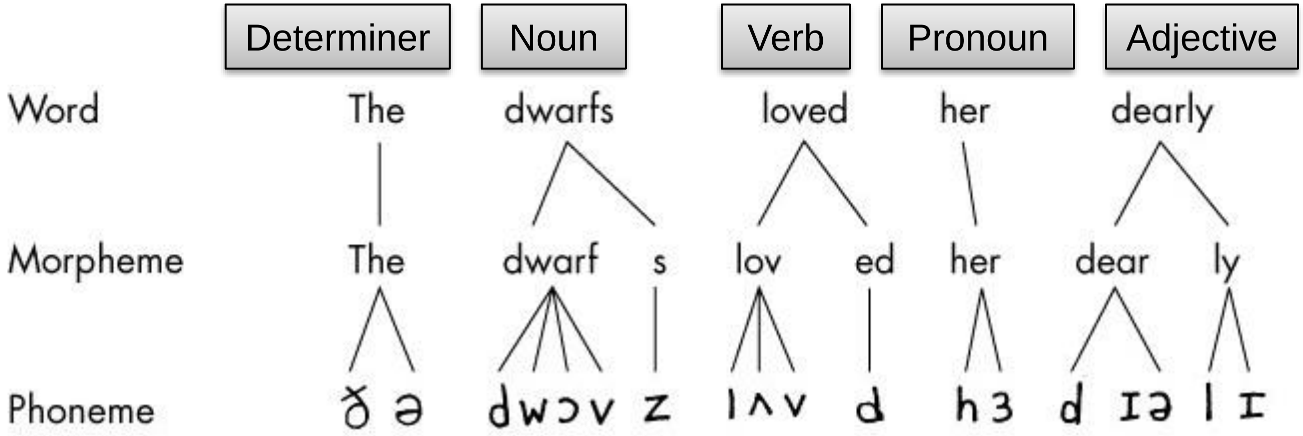
Menti.com
2848 6424

- **Syntax** refers to the way words are arranged together
- "Syntax is the study of the regularities and constraints of word order and phrase structure"
  (Manning & Schütze, 2003, p. 93)
- There is an infinite number of ways in which words can be arranged together to form sentences
- Yet, we can understand sentences we have never heard or read before

POS Tagging

Menti.com
2848 6424

- The process of assigning **a part of speech** or **lexical class marker** to each word in a corpus
- The input to a tagging algorithm is a sequence of words and a tagset, and the output is a sequence of tags, a single best tag for each word



(c) David Groome, 2006

Parts of Speech

Menti.com
2848 6424

- In English we traditionally have 8 parts of speech

N	Noun	*chair, bandwidth, pacing*
V	Verb	*study, debate, munch*
ADJ	Adjective	*purple, tall, ridiculous*
ADV	Adverb	*unfortunately, slowly*
P	Preposition	*of, by, to*
PRO	Pronoun	*I, me, mine*
DET	Determiner	*the, a, that, those*
INTJ	Interjection	*oh!, m-hm, huh?*

Penn Treebank Tagset

Menti.com
2848 6424

1. CC Coord. conjunc.
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Prep./subord. conj.
7. JJ Adject.
8. JJR Adject., comp.
9. JJS Adject., superl.
10. LS List item marker
11. MD Modal
12. NN Noun, sing. or mass
13. NNS Noun, plural
14. NNP Proper noun, sing.
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PP$ Poss. pronoun
20. RB Adverb
21. RBR Adverb, comp.
22. RBS Adverb, superl.
23. RP Particle
24. SYM Symbol
25. TO to
26. UH Interjection
27. VB V, base form
28. VBD V, past tense
29. VBG V, gerund/pres. part.
30. VBN V, past part.
31. VBP V, non-3rd ps. sing. pres.
32. VBZ V, 3rd ps. sing. pres.
33. WDT wh-det.
34. WP wh-pronoun
35. WP$ Poss. wh-pronoun
36. WRB wh-adverb
37. # Pound sign
38. $ Dollar sign
39. . Sent.-final punct.
40. , Comma
41. : Colon, semi-colon
42. (L. bracket char.
43.) R. bracket char.
44. “ Straight dbl. quote
45. ‘ L. open sngl. quote
46. “ L. open dbl. quote
47. ’ R. close sngl. quote
48. ” R. close dbl. quote

Language	**Tagset Size**
English	139
Czech	970
Estonian	476
Hungarian	401
Romanian	486
Slovene	1033

(Hajič, 2000)

An Example

Menti.com
2848 6424

WORD	LEMMA	TAG
the	the	+DET
host	host	+NOUN
kissed	kiss	+VPAST
the	the	+DET
friend	friend	+NOUN
on	on	+PREP
the	the	+DET
cheek	cheek	+NOUN

Ambiguities

Menti.com
2848 6424

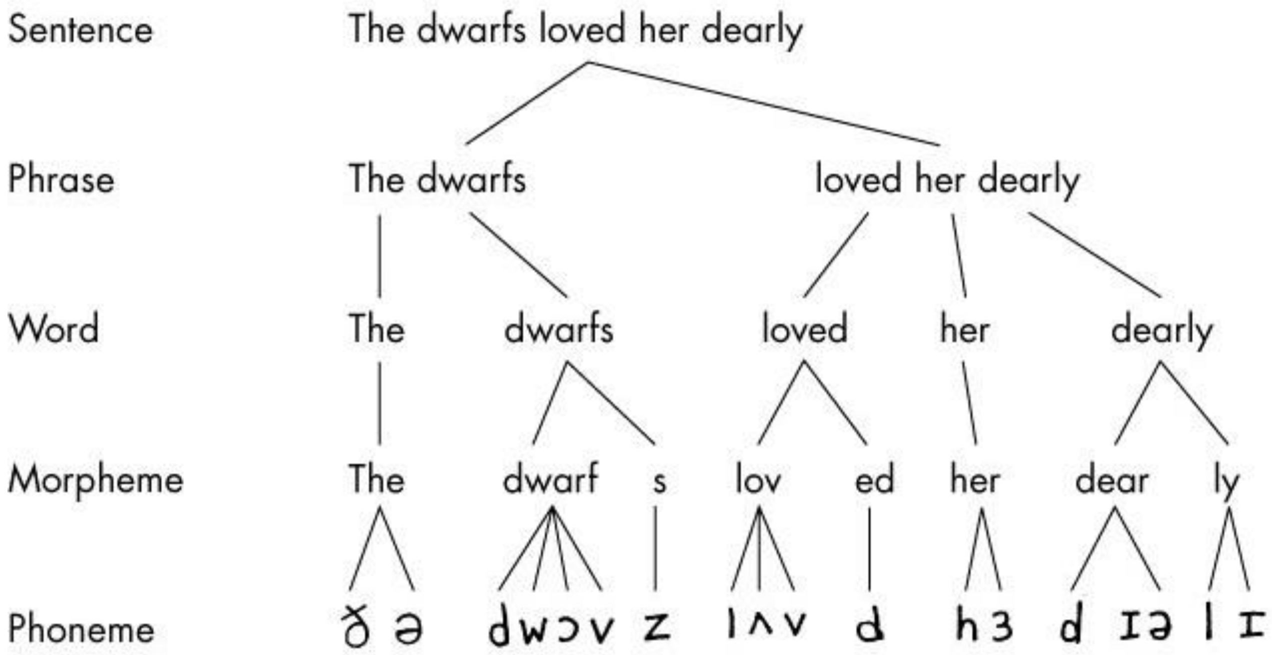
- POS Tagging is a disambiguation task
  - Words are ambiguous—have more than one possible part-of-speech
  - The word “book”:
    - book that flight: verb
    - hand me that book: noun
  - The word “that”:
    - Does that flight serve dinner? : determiner
    - I thought that your flight was earlier: complementizer

- **POS Tagging**: resolves ambiguities, choosing the proper tag for the context
  - Baseline: Most Frequent Class (accuracy 92.34% [Jurafsky & Martin])
  - Outdated: Rule-based tagging, probabilistic tagging
  - State of the art: Neural approaches, accuracy ~ 98%

Parsing

Menti.com
2848 6424

- The process of determining the grammatical structure with respect to a given grammar.



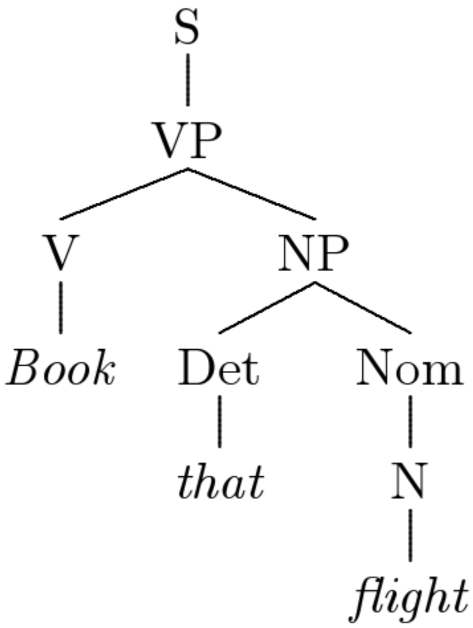
(c) David Groome, 2006

Alternative representations

Menti.com
2848 6424

- Bracketed notation:
  [$_S$ [$_{NP}$ [$_{Det}$ the] [$_N$ dog]] [$_{VP}$ [$_V$ ate] [$_{NP}$ [$_{Det}$ a] [$_N$ cookie]]]]
- Parenthesized notation:
  (S
  (NP
  (Det the)
  (N dog))
  (VP
  (V ate)
  (NP
  (Det a)
  (N cookie))))

Parse Tree:



Syntactic Ambiguity

Menti.com
2848 6424

- If you love money problems show up
  - If you love, money problems show up.
  - If you love money, problems show up.
  - If you love money problems, show up.
- “I made her duck.”
- “We're eating grandpa!” vs. "We're eating, grandpa!"
- “Weil er drei Monate verfallene Medikamente nahm, ...”

- Different interpretations are mainly caused by syntactic ambiguity.

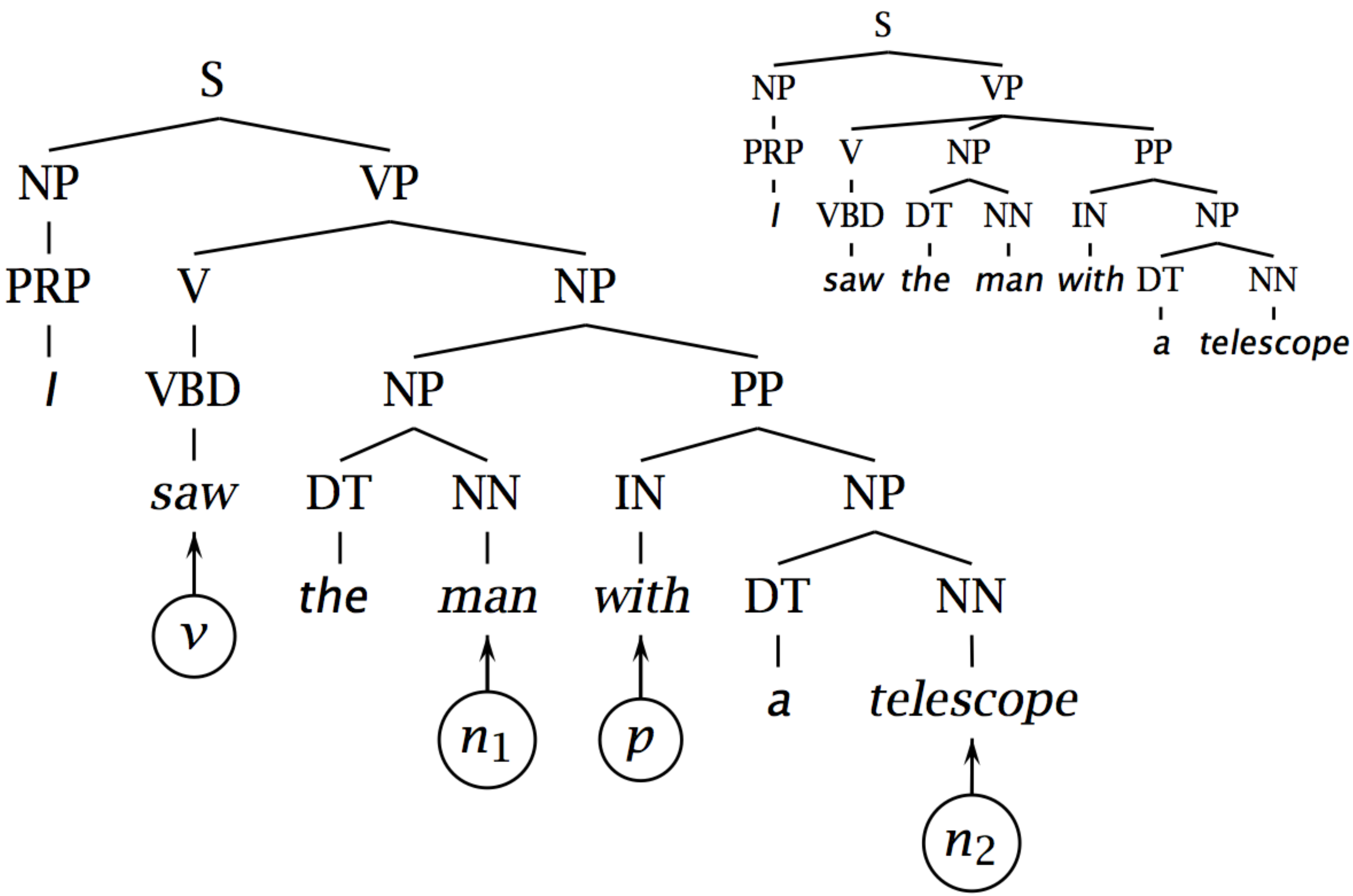
Syntactic Ambiguities:
Two Possible Parsing Possibilities

Menti.com
2848 6424

“I saw the man with a telescope.”

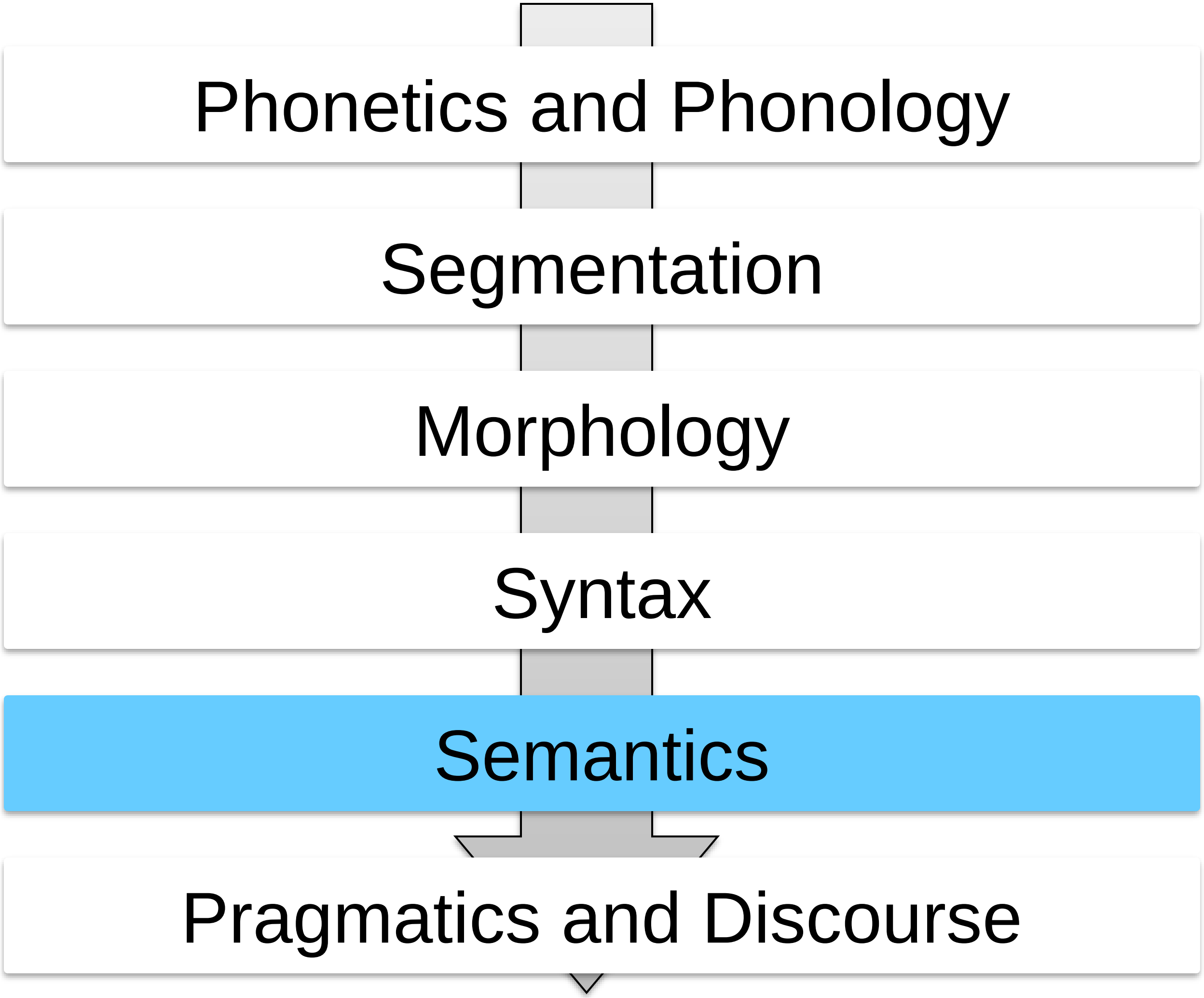
Syntactic Ambiguities:
Two Possible Parsing Possibilities

Menti.com
2848 6424



Analysis Levels in Language Understanding

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

Syntax

Semantics

Pragmatics and Discourse

Definition

Menti.com
2848 6424

Semantics:

- Study of the meaning of **words, phrases, sentences, or documents**

Lexical Semantics

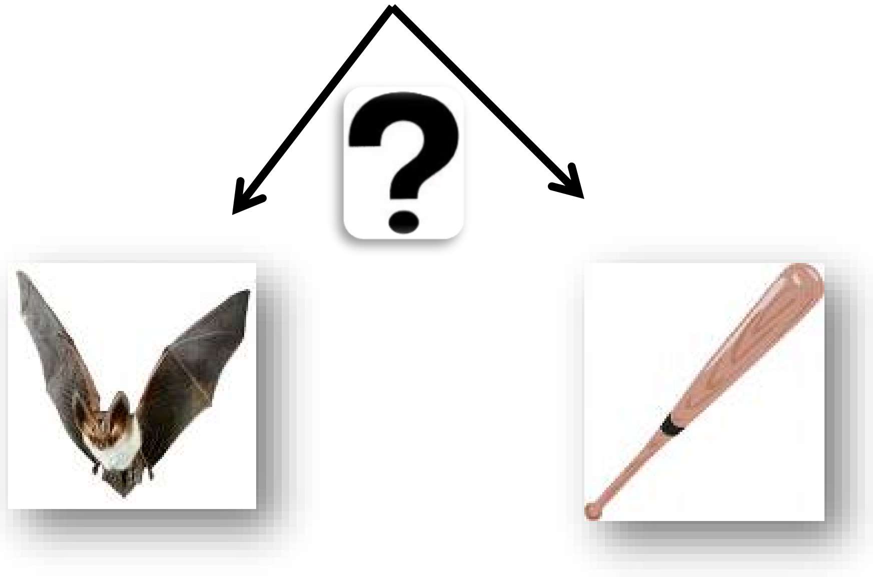
- Study of the meaning of lexical units, i.e. **words**.

Lexical Ambiguity

Menti.com
2848 6424

He hit the ball with the bat.

Chuck Norris can hit a bat with a ball.



- Different interpretations are caused by lexical ambiguity.

Analysis Levels in Language Understanding

Menti.com
2848 6424

- Phonetics and Phonology
- Segmentation
- Morphology
- Syntax
- Semantics
- **Pragmatics and Discourse**

Pragmatics

Menti.com
2848 6424

What is the purpose of an utterance?

“I **never** said she stole my money"	I simply didn't ever say it.
▪ “**I** never said she stole my money”	Someone else said it, but I didn't.
▪ “I never **said** she stole my money”	I might have implied it in some way, but I never explicitly said it.
▪ “I never said **she** stole my money”	I said someone took it; I didn't say it was she.
▪ “I never said she **stole** my money”	I just said she probably borrowed it.
▪ “I never said she stole **my** money”	I said she stole someone else's money.
▪ “I never said she stole my **money**”	I said she stole something of mine, but not my money.

Example from Wikipedia

Pragmatics

Menti.com
2848 6424

What is the purpose of an utterance?

Utterance: “Is it cold in here or is it just me?

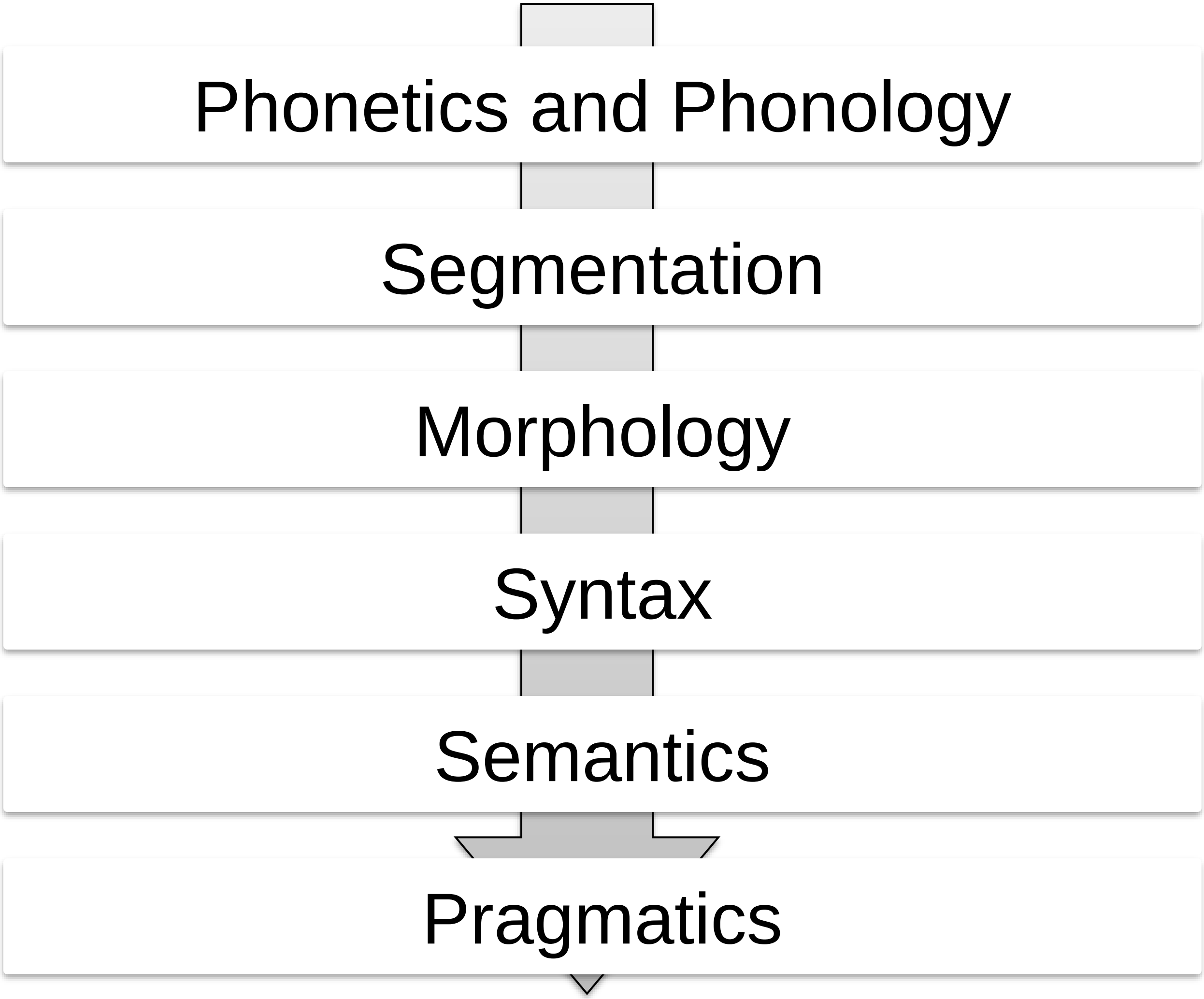
Intended meaning: “Please close the window!”

Utterance: “Oh, great! Another meeting.”

Intended meaning: The speaker likely means the opposite of what they are literally saying—meetings might be something they dislike, despite the positive tone.

Summary – Linguistic Analysis Levels

Menti.com
2848 6424



Phonetics and Phonology

Segmentation

Morphology

Syntax

Semantics

Pragmatics

Summary – Linguistic Analysis Levels

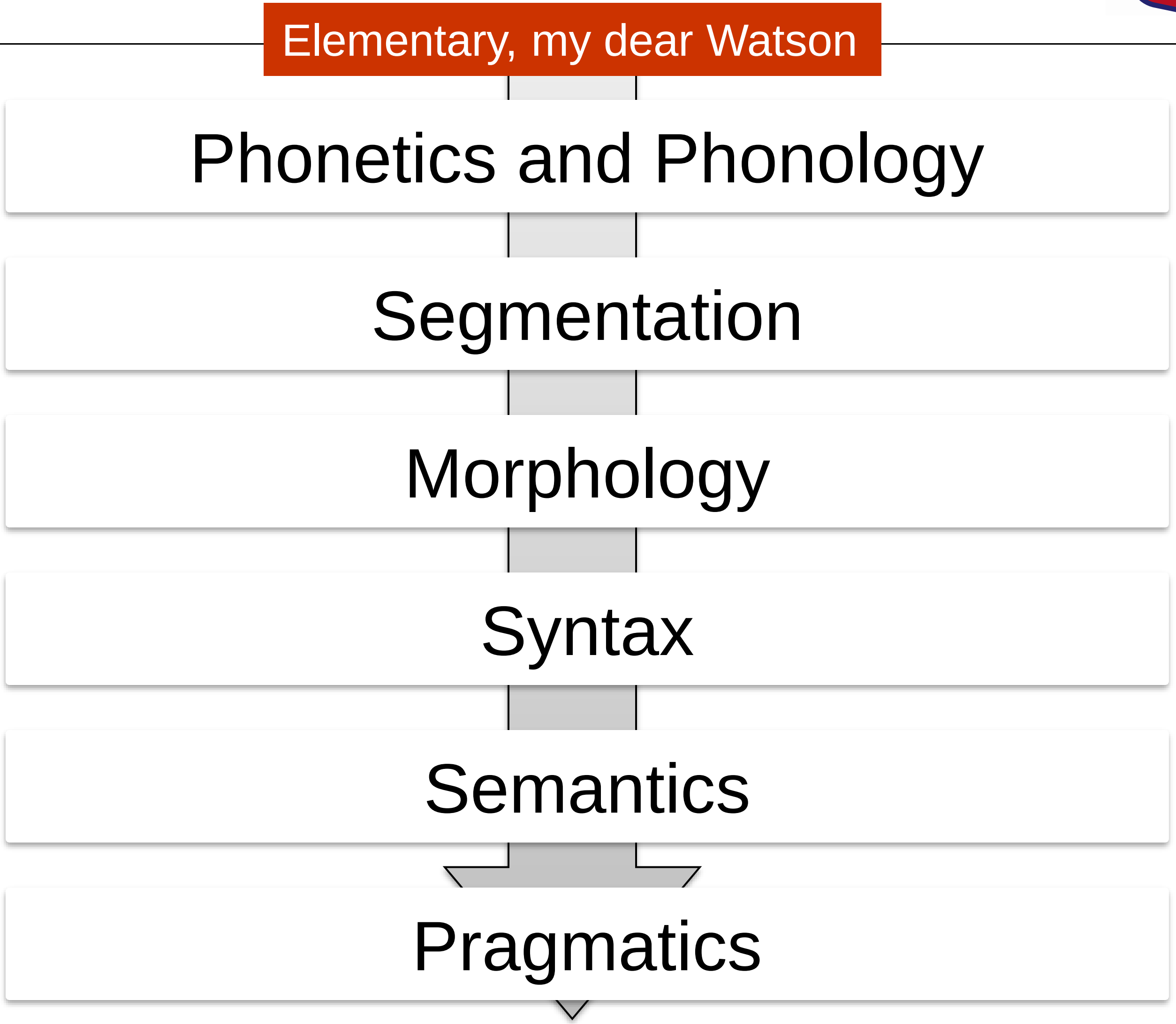
Menti.com
2848 6424

Elementary, my dear Watson

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson



- Phonetics and Phonology
- Segmentation
- Morphology
- Syntax
- Semantics
- Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

Segmentation

Morphology

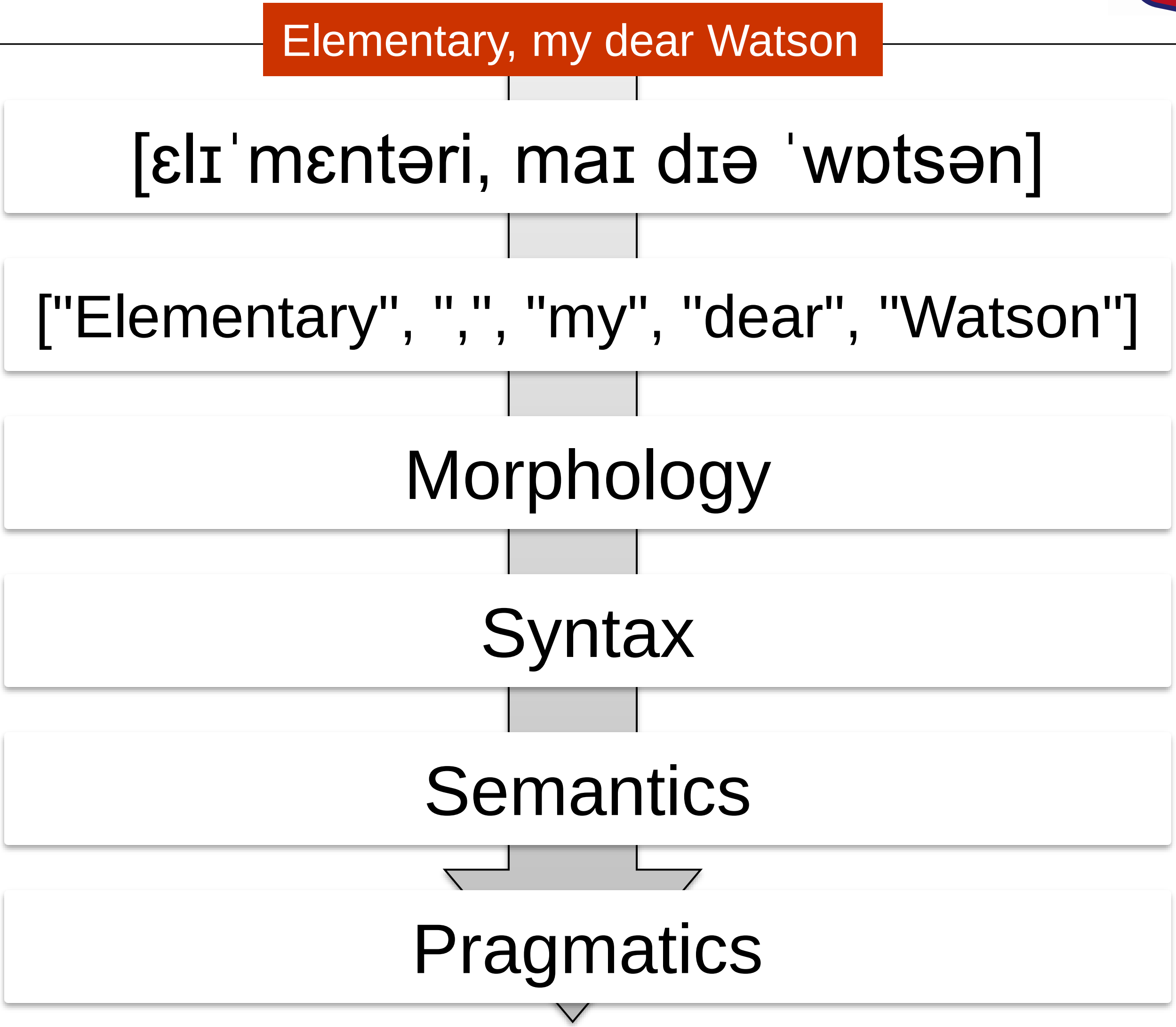
Syntax

Semantics

Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424



Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

["Elementary", ",", "my", "dear", "Watson"]

Morphology

Syntax

Semantics

Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

["Elementary", ",", "my", "dear", "Watson"]

Base: Element, Suffix: -ary

Syntax

Semantics

Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

["Elementary", ",", "my", "dear", "Watson"]

Base: Element, Suffix: -ary

ADJ, PRP$ ADJ NNP

Semantics

Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

["Elementary", ",", "my", "dear", "Watson"]

Base: Element, Suffix: -ary

ADJ, PRP$ ADJ NNP

Watson: Dr. John H. Watson (not IBM)

Pragmatics

Summary – Linguistic Analysis Levels

Menti.com
2848 6424

Elementary, my dear Watson

[ɛlɪˈmɛntəri, maɪ dɪə ˈwɒtsən]

["Elementary", ",", "my", "dear", "Watson"]

Base: Element, Suffix: -ary

ADJ, PRP$ ADJ NNP

Watson: Dr. John H. Watson (not IBM)

"You are so stupid…"

Take-Home-Messages

Menti.com
2848 6424

- Natural language processing is an interesting topic ☺
- There are a lot of challenges ☹

- Typical preprocessing steps:
  - Tokenization for splitting texts into tokens
  - Stemming / Lemmatization to normalize tokens
  - PoS-Tagging and parsing analyze syntactic features
    - PoS-tags roughly represent word classes
    - Phrases group words to function as a single unit

- Ambiguity in language makes analysis a hard problem

Next Lecture

Menti.com
2848 6424

Text Classification