

# Statistical Natural Language Processing

## Part II: Basics of NLP

Henning Wachsmuth

<https://ai.uni-hannover.de>

# Learning Objectives

## Concepts

- Basic concepts from linguistics
- Standard evaluation measures in NLP
- The most relevant basics from statistics

## Methods

- Fundamental techniques in NLP
- Selection of the right evaluation measure for a task
- The study of hypotheses with significance tests

## Notice

- Some concepts and methods are reviewed briefly here only.
- For more details, see the content of my bachelor's lecture:  
[Introduction to Natural Language Processing](#)

# Outline of the Course

## I. Overview

## II. Basics of NLP (video only)

- Introduction
- Linguistics
- Fundamental NLP Techniques
- Evaluation Measures
- Empirical Experiments
- Conclusion

## III. Basics of Statistical NLP

## IV. Representation Learning

## V. NLP using Clustering

## VI. NLP using Classification and Regression

## VII. NLP using Sequence Labeling

## VIII. NLP using Neural Networks

## IX. NLP using Transformers

## X. Practical Issues

# Introduction

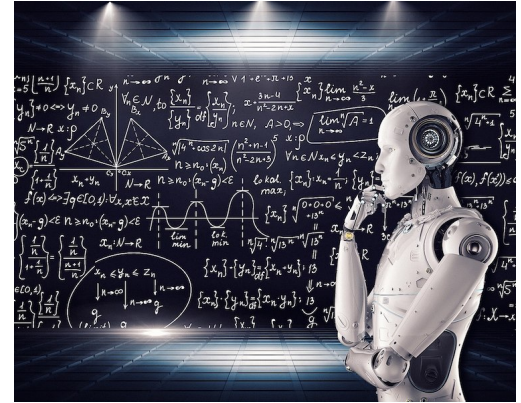
# Natural Language Processing (NLP) (recap)

## Natural language processing

- The study of computational methods for understanding and generating human-readable text (or speech)

We mostly speak about text only in this course.

- The goal is to decode structured information from language, or to encode it in language.
- NLP is a subfield of AI, and one part of computational linguistics.



<https://wikimedia.org>

## Computational linguistics

- Roughly, the intersection of computer science and linguistics
- **Technologies** for natural language processing
- **Models** to explain linguistic phenomena, using knowledge or statistics

## Linguistics

- The study of natural language(s) in terms of form, meaning, and context

# Natural Language Processing (NLP)

## Analysis and Synthesis

### Types of NLP tasks

- **Analysis.** The inference of structured information from text (decoding)  
*Analysis tasks are referred to as *natural language understanding (NLU)*.*
- **Synthesis.** The generation of text from structured information or from other text (encoding)  
*Synthesis tasks are referred to as *natural language generation (NLG)*.*

### Selected analysis tasks

- Token and sentence splitting
- Syntactic parsing
- Entity recognition
- Reference resolution
- Relation extraction
- Topic detection
- Sentiment analysis

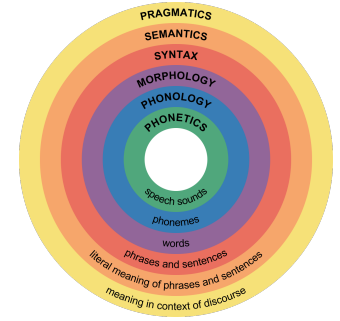
### Selected synthesis tasks

- Grammatical error correction
- Sentence generation
- Discourse composition
- Summarization
- Text style transfer
- Cluster labeling
- Lexicon creation

# Basics of NLP

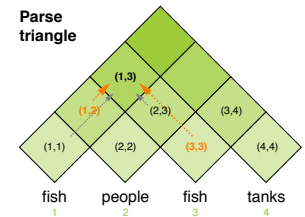
## Linguistic units and levels

- The encoding and decoding of information builds on various concepts across all linguistic levels.
- This requires studying text units and their structure, understanding their semantics, and their pragmatics.



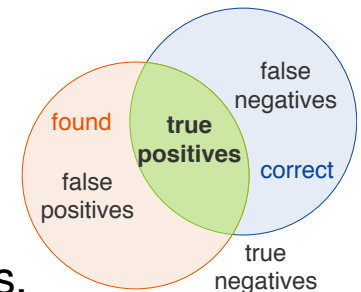
## Fundamental NLP techniques

- NLP combines various analysis and synthesis steps.
- Traditional NLP methods rely on rules, enhanced by statistics such as frequencies and probabilities.



## Evaluation and experiments

- The quality of NLP methods needs to be measured, especially their effectiveness.
- Methods are evaluated empirically on test data and compared to alternative methods using statistical tests.



# Basics of NLP

## Development and Evaluation

### Need for data

- NLP methods tackle specific analysis or synthesis tasks.
- To this end, they operationalize expert rules and/or statistical patterns.
- Rules and patterns are derived from analyses of training data.

### Need for evaluation

- The output of NLP methods is rarely free of errors due to the ambiguity of language.
- Thus, they are evaluated empirically on test data.
- The *effectiveness* of methods is quantified with measures such as accuracy.



<https://pixabay.com>

### Need for comparison

- It is unclear per se how good a measured value is in a given task.
- Methods are thus compared to other methods, so called *baselines*.

# Linguistics

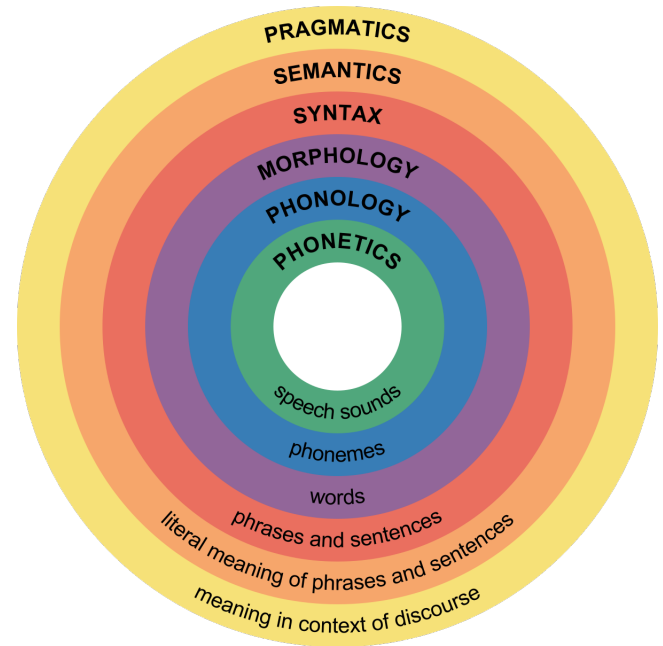
# Linguistics

## Linguistics

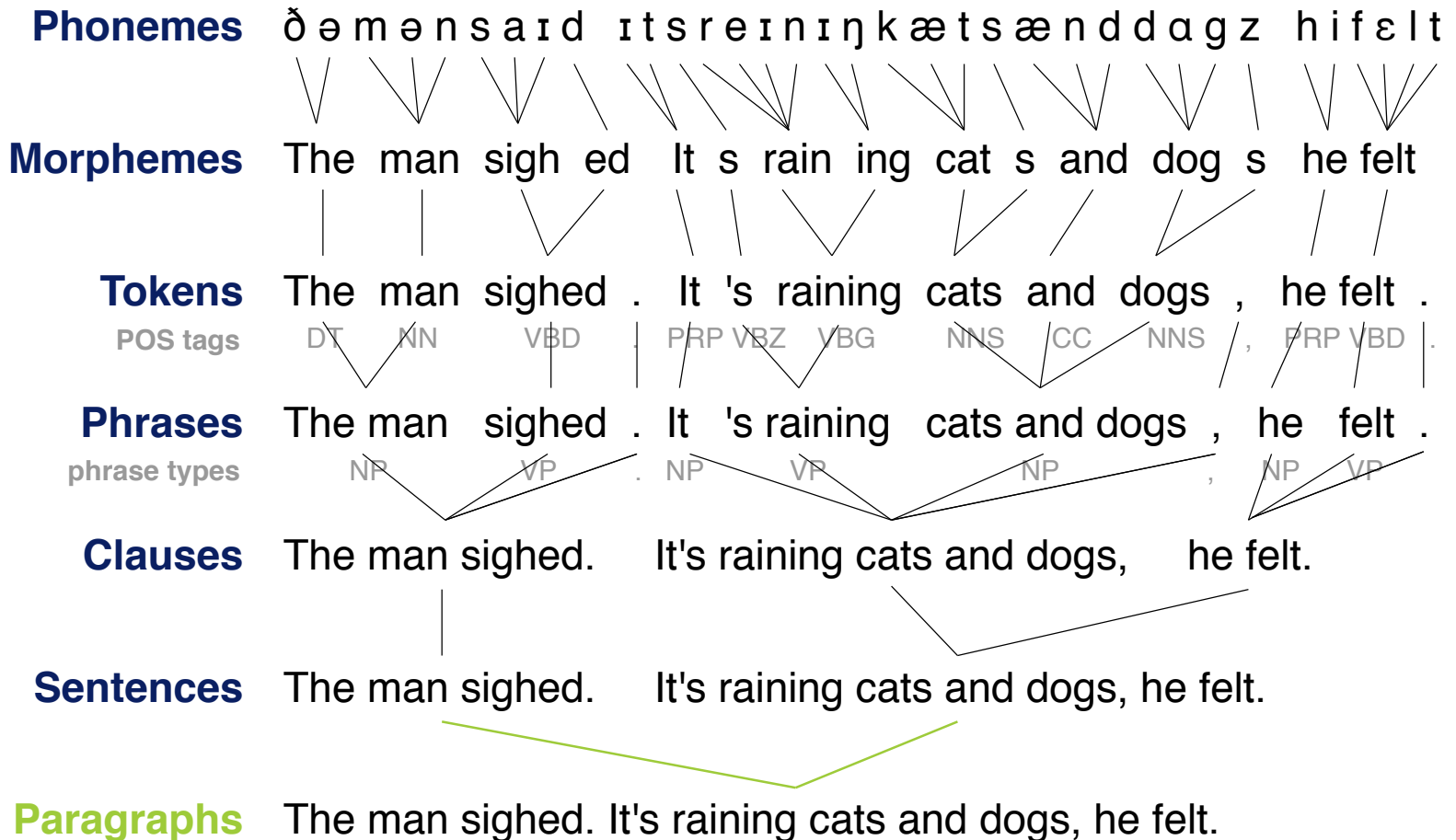
- The study of natural language(s) in terms of form, meaning, and context

## Linguistic Levels

- **Phonetics**. Physical aspects of speech sounds
- **Phonology**. Linguistic sounds of a particular language
- **Morphology**. Senseful components of words
- **Syntax**. Structural relationships between words, usually in a sentence
- **Semantics**. Meaning of single words and compositions of words
- **Discourse**. Composition of linguistic units larger than a sentence
- **Pragmatics**. Use of language to accomplish certain goals



# Linguistics



# Linguistics

## Main Morphological Concepts

### Word

- The smallest unit of language that is to be uttered in isolation

“cats” and “ran” in “cats ran.”

### Lemma

- The dictionary form of a word

“cat” for “cats”

“run” for “ran”

### Stem

- The part of a word that never changes

“cat” for “cats”

“ran” for “ran”

### Token

- The smallest text unit in NLP: A word, number, symbol, or similar  
Whitespaces are usually not considered as tokens.

“cats”, “ran”, and “.” in “cats ran.”

# Linguistics

## Main Syntactic Concepts

### Part-of-speech (POS)

- The lexical category (or word class) of a word
- **Abstract classes.** Nouns, verbs, adjectives, adverbs, prepositions, ...
- **POS tags.** NN (single nouns), NNS (plural n.'s), NNP (proper n.'s), ...

### Phrases

- A contiguous sequence of words, functioning as a single meaning unit
- Phrases often contain nested phrases.
- **Types.** Noun phrase (NP), verb phrase (VP), prepositional phrase (PP)  
Sometimes also adjectival phrase (AP) and adverbial phrase (AdvP)

### Clause

- The smallest grammatical unit that can express a complete proposition
- **Types.** Main clause, subordinate clause

### Sentence

- A grammatically independent linguistic unit with one or more words

# Linguistics

## Main Semantic Concepts

### Two types of semantics

- **Lexical.** The meaning of words and multi-word expressions
- **Compositional.** The meaning of the composition of words

### Entities

- An object from the real world
- **Named entities.** Persons, locations, organizations, products, ...

“Prof. Dr. Henning Wachsmuth”

“Hannover”

“Uni Hannover”

- **Numeric entities.** Values, quantities, ranges, periods, dates, ...

“in this year”

“2024-10-17”

“\$ 100 000”

“762-12377”

### Relations

- **Semantic.** Relations between entities, e.g., *person lives in location*
- **Temporal.** Relations describing courses of events, e.g., *after <A>, <B>*

# Linguistics

## Main Discourse and Pragmatics Concepts

### Discourse (structure)

- Linguistic utterances larger than a sentence, e.g., paragraphs or articles
- **Discourse segment.** Linguistic building block within a discourse
- **Coherence relation.** Semantic or pragmatic relations between segments

### Coreference

- Two or more expressions in a text that refer to the same thing
- **Types.** Pronouns, coreferring noun phrases, ...

“Apple Inc. is based in the US. The company is called Apple. They make hardware.

### Speech acts

- Linguistic utterances with a performative function

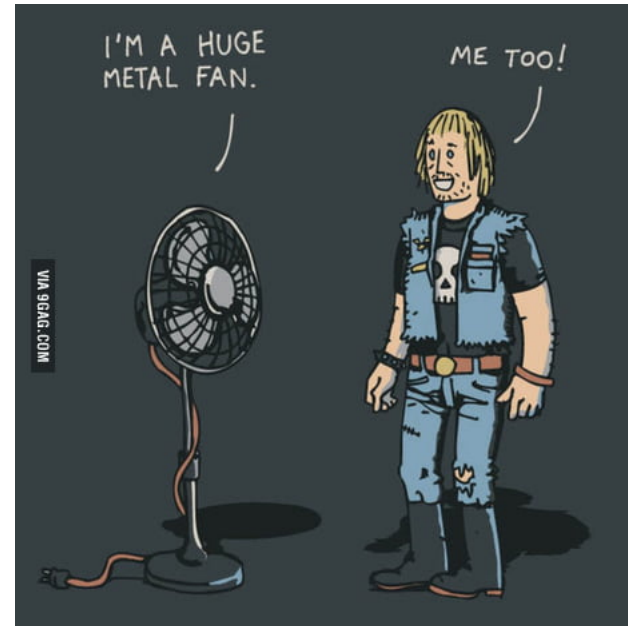
### Presupposition and implicature

- **Presupposition.** Linguistic utterances assume things.
- **Implicature.** Linguistic utterances suggest things.

# Challenges in Language Understanding

## Ambiguity in natural language

- **Phonetic.** “wreck a nice beach”
- **Word sense.** “I went to the bank.”
- **Part of speech.** “I made her duck.”
- **Attachment.** “I saw a man with a telescope.”
- **Coordination.** “If you love money problems show up.”
- **Quantifier scope.** “I didn’t buy a car.”
- **Speech act.** “Have you emptied the dishwasher?”



## Selected other challenges

- **World knowledge.** “Putin’s view of Ukraine is wrong”
- **Domain dependency.** “Read the book!”
- **Language dependency.** “Bad”

# Fundamental NLP Techniques

# Fundamental NLP Techniques

## Recap: Rule-based NLP

- Based mainly on manually defined rules that encode expert knowledge
- Knowledge includes rules, lexicons, grammars, and similar.

## Statistics in rule-based NLP

- Some techniques employ statistics to make decisions or to weigh rules.
- As such, they reflect the transition from rule-based to statistical NLP.

## Selected techniques

- **Decision trees.** Apply nested if-then-else rules to a text.
- **Lexicon matching.** Match text spans with terms from a lexicon.
- **Regular expressions.** Extract text spans that follow sequential patterns.
- **Probabilistic parsing.** Infer hierarchical structures of text spans.
- **Template-based generation.** Create texts by filling predefined slots.
- **Language modeling.** Generate sequences of words and other tokens.

# Fundamental NLP Techniques

## Decision Trees

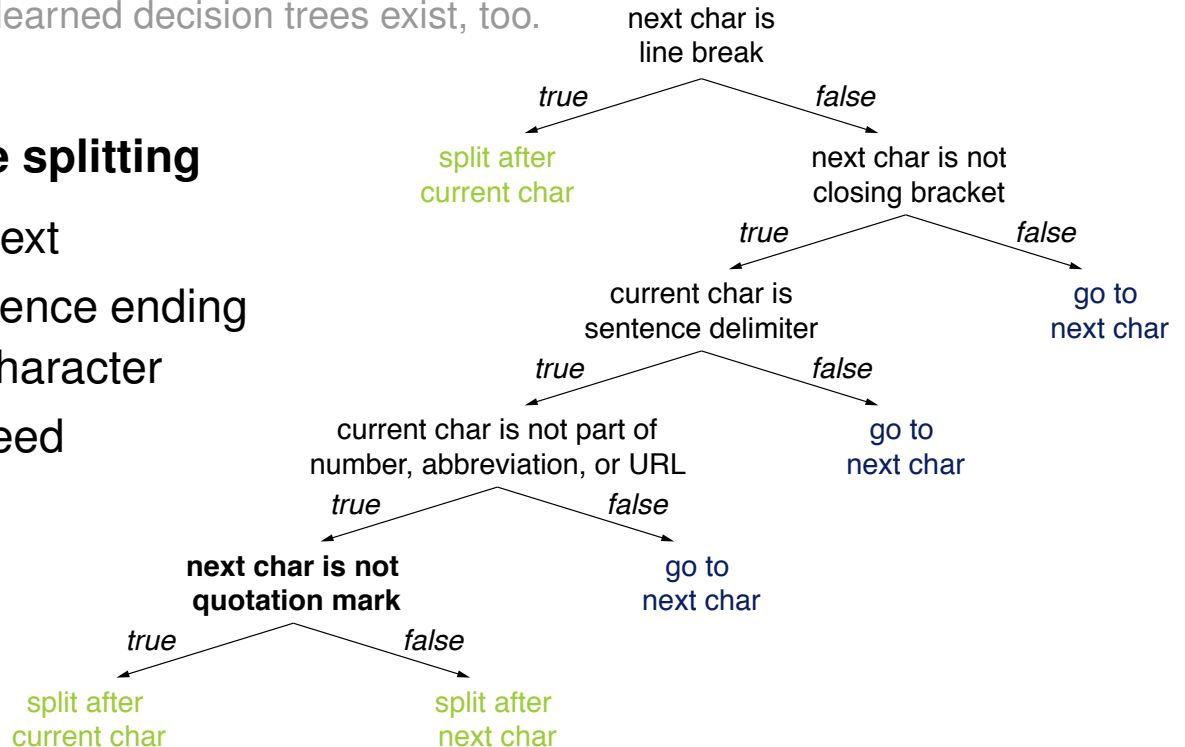
### Decision trees

- The representation of a series of if-then-else decision rules
- Inner nodes are decision criteria, leaves the final outcomes in a task.
- **Hand-crafted.** Rules are composed using expert knowledge.

Notice: Machine-learned decision trees exist, too.

### Example: Sentence splitting

- Given a plain text
- Check for sentence ending character by character
- Split and proceed



# Fundamental NLP Techniques

## Lexicon Matching

### Several types of lexicons

- **Terms.** Word lists, language lexicons, vocabularies
- **+ Definitions.** Dictionaries, glossaries, thesauri
- **+ Structured information.** Gazetteers, frequency lists, confidence lists

### Use cases of lexicons

- A given lexicon can be used to find all term occurrences in a text.
- The existence of a given term in a lexicon can be checked.
- The density or distribution of a vocabulary in a text can be measured.

### Example: Attribute extraction

- Given a training set with annotated attributes
- Compute confidence of each term, i.e., how often it is annotated as attribute
- Consider terms with confidence above some threshold as attributes

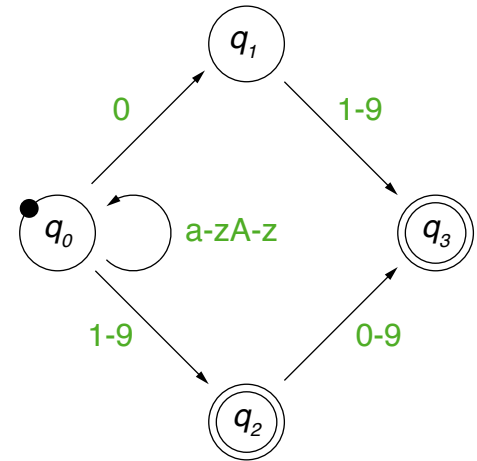
Attribute	Confidence
minibar	1.00
towels	0.97
wi-fi	0.83
front desk	0.74
alcohol	0.50
buffet	0.21
people	0.01

# Fundamental NLP Techniques

## Regular Expressions

### Regular expression (regex)

- A representation of a regular grammar
- Combines characters and meta-characters to generalize over language structures
- Used in NLP mainly to match text spans that follow clear sequential patterns



### Types of patterns in regexes

- **Disjunctions.** Alternatives, such as `([Ww]oodchuck | [Gg]roundhog)`
- **Negation+choice.** Restrictions and wildcards, such as `[^A-Z]` or `19..`
- **Repetitions.** Parts that are optional and/or may appear multiple times, such as `woo(oo)?dchuck`, `woo(oo)*dchuck`, or `woo(oo)+dchuck`

### Example: German date extraction

```
(0?[1-9] | [10-31])\. (0?[1-9] | [10-12])\. (19|20) [0-9] [0-9]
```

# Fundamental NLP Techniques

## Probabilistic Parsing

### Probabilistic context-free grammar (PCFG)

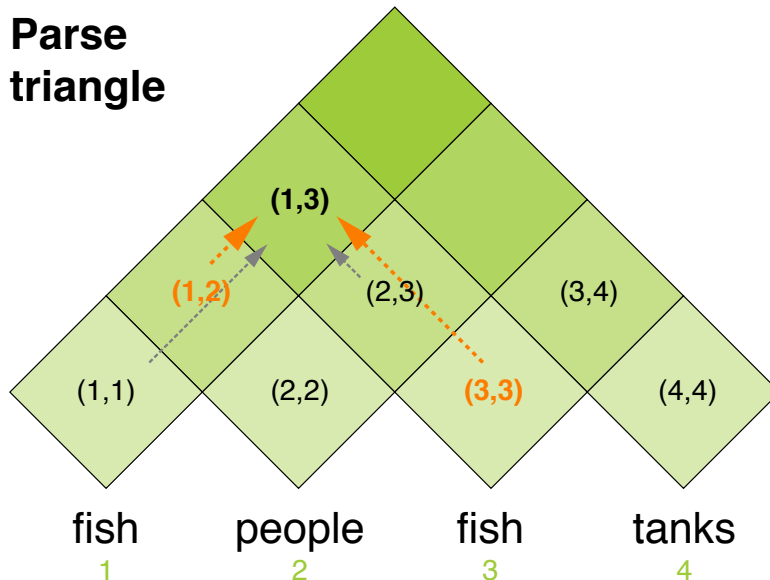
- A CFG where each rule is assigned a probability
- Used in NLP mainly to parse sentence structure
- The goal is to find the most likely parse tree

Rule	Prob.
$S \rightarrow NP VP$	1.0
$VP \rightarrow V NP$	0.6
$VP \rightarrow V NP PP$	0.4
...	...
$V \rightarrow \text{fish}$	0.6
$V \rightarrow \text{tanks}$	0.3

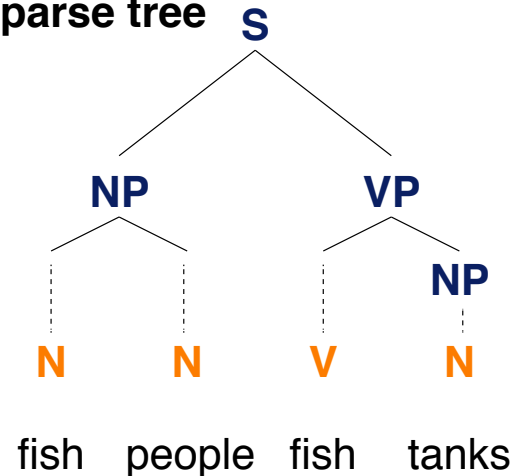
### Example: Constituency parsing

- Use dynamic programming to iteratively compute the most likely tree

Parse triangle



Most likely parse tree



# Fundamental NLP Techniques

## Template-based Generation

### Template-based Generation

- The generation of a sentence (or similar) by filling slots of a predefined sentence template with specific information
- If any, phrasing is done only to account for grammar and coherence.

Examples: Change singular to plural, add discourse markers, capitalize, ...

### Example: Claim generation

- **Template.** “I am <stance> <issue>, because <reason>.”

**Issue.** Death penalty

**Stance.** Pro

**Reason.** The death penalty kills people



“I am pro death penalty, because the death penalty kills people.”

### Origin of templates

- Traditionally, the templates are manually defined by experts.
- An alternative is to derive them from a given corpus.

# Fundamental NLP Techniques

## Language Modeling

### Language model

- A probability distribution over a sequence of words

Assigns a probability  $P(w_1, \dots, w_m)$  to each sequence of words  $w_1, \dots, w_m$  for any  $m$

- **$n$ -gram model.** Approximates the probability of  $m$  words for some  $n$  as:

$$P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1}) \approx \prod_{i=1}^m P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$$

- The probabilities can be derived from all word sequences in a corpus.

### Language models in generation

- Given an  $n$ -gram, the most likely words following it can be computed.
- **Example.** Next two words after “fish” based on a 2-gram model

$$\begin{array}{llll} P(\text{fish}|\text{fish}) = 0.2 & P(\text{fish}|\text{people}) = 0.6 & \rightarrow & P(\text{people people}|\text{fish}) = 0.32 \quad \dots \\ P(\text{people}|\text{fish}) = 0.8 & P(\text{people}|\text{people}) = 0.4 & & P(\text{people fish}|\text{fish}) = 0.48 \quad \dots \end{array}$$

# Fundamental NLP Techniques

## Terminology

### Terms in SNLP

- **Task.** A specific problem with a defined input and desired output  
Examples: Classification of sentiment polarity, generation of a text summary, ...
- **Technique.** A general way of how to analyze and/or synthesize a text  
Examples: Feature-based clustering, transformer-based text generation, ...
- **Algorithm.** A specific implementation of a technique  
Examples:  $k$ -means, BART, ...
- **Model.** The configuration of an algorithm resulting from training  
Examples: 5-means trained on data xxx, BART fine-tuned on data yyy
- **Approach.** A computational method using model(s) to tackle a task  
Example: A method that summarizes argumentative text using fine-tuned BART
- **Application.** A technology that tackles a real-world problem using NLP  
Example: Google Assistant

### Notice

- Informally, the terms method, algorithm, model, and approach are often used more or less interchangeably.

# Fundamental NLP Techniques

## NLP Processes

### From single tasks to processes

- Most NLP applications aim at combinations of different types of output.
- This means that several analysis or synthesis steps may be needed.

### Ways to realize NLP processes

- **Pipelines.** Sequentially apply a set of methods to a text, such that the output of one method is the input to the next.
- **Joint models.** Perform multiple analysis/synthesis steps simultaneously.
- **Neural models.** Neural networks often operate on the raw input text.

In any way, some kind of sequential pipeline is used for most NLP processes.

### Example: Information extraction (IE)

- The mining of entities, their attributes, and their relations from text
- Usually, IE requires a pipeline of several analysis steps.
- The output is structured information, e.g., for the use in databases.

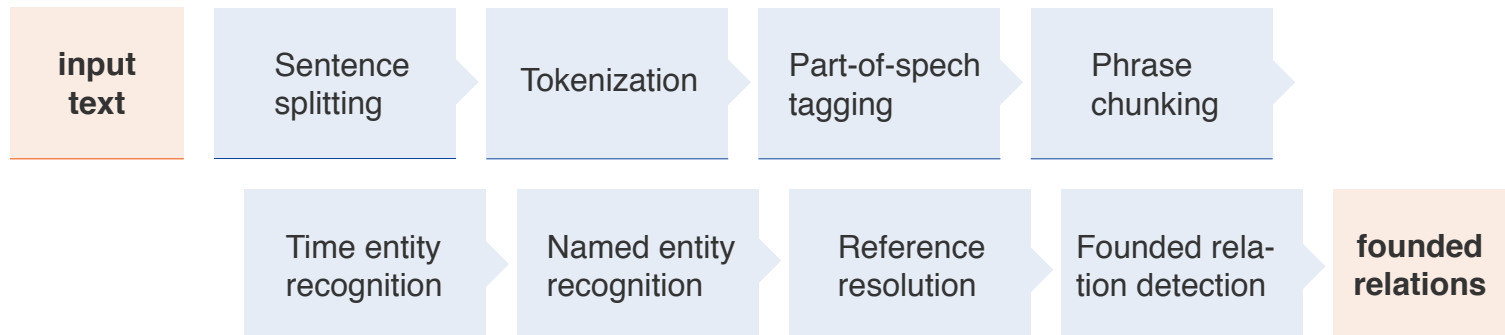
# Fundamental NLP Techniques

## Example: Information Extraction

### Example: Extraction of founding dates



### Text analysis pipeline for this example



# Evaluation Measures

# Evaluation Measures

## Evaluation measures in NLP

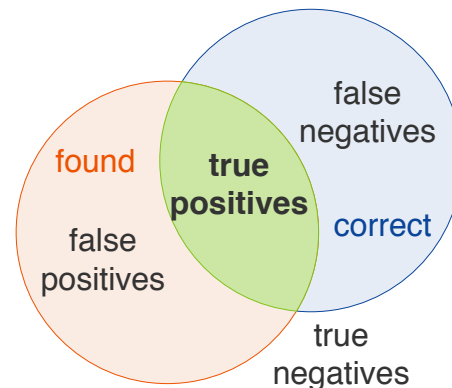
- A measure to quantify a method's quality on a given task and corpus
- Methods can be ranked with respect to an evaluation measure.
- Quality is mostly assessed in terms of *effectiveness*, that is, the extent to which the output information of a method is correct

## Measuring effectiveness

- How to adequately measure effectiveness, depends on the task.
- **Analysis.** The output of a method is compared to the ground truth.
- **Synthesis.** Usually, not only one correct output exists.

## Instance types in (analysis) tasks

- **True positive (TP).** Correctly found
- **True negative (TN).** Correctly not found
- **False negative (FN).** Mistakenly not found
- **False positive (FP).** Mistakenly found



# Evaluation Measures

## Accuracy

### Accuracy

- The accuracy  $A$  is a measure of the correctness of a method.
- For  $m = 2$  classes, accuracy is the ratio of positives under all instances.

$$A_{binary} := \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

- For  $k > 2$  classes, accuracy is simply the ratio of true positives.

$$A_{multi} := \frac{|TP_1| + \dots + |TP_k|}{|TP_1| + \dots + |TP_k| + |FP_1| + \dots + |FP_k|}$$

### When to use accuracy?

- Accuracy is often adequate when all classes are of similar importance.  
Examples: Sentiment analysis, part-of-speech tagging, ...

# Evaluation Measures

## Precision and Recall

### Precision

- The precision  $P$  is a measure of the exactness of an approach.
- $P$  answers: How many of the found instances are correct?

$$P := \frac{|TP|}{|TP| + |FP|}$$



### Recall

- The recall  $R$  is a measure of the completeness of an approach.
- $R$  answers: How many of the correct instances have been found?

$$R := \frac{|TP|}{|TP| + |FN|}$$



### When to use precision and recall?

- Together, they are adequate if the focus is on the positive instances.  
Examples: Named entity recognition, plagiarism detection, ...

# Evaluation Measures

## $F_1$ -score and Averaging

### **$F_1$ -score** (aka $F_1$ -measure)

- The  $F_1$ -score is the harmonic mean of precision and recall.
- $F_1$  favors balanced over imbalanced precision and recall values.

$$F_1 := \frac{2 \cdot P \cdot R}{P + R}$$

### **Multi-class precision** (recall and $F_1$ -score analogous)

- In general, each class in a multi-class task can be evaluated binarily.
- Overall results are obtained with micro- or macro-averaging.
- **Micro-averaging.** Take into account the number of instances per class:

$$P_{micro} := \frac{|TP_1| + \dots + |TP_k|}{|TP_1| + \dots + |TP_k| + |FP_1| + \dots + |FP_k|}$$

- **Macro-averaging.** Compute the mean result over all classes:

$$P_{macro} := \frac{P_1 + \dots + P_k}{k}$$

# Effectiveness Measures for Numerical Predictions

## Evaluating real-valued predictions

- It is unlikely to predict exact real values in many tasks.
- In such cases, the focus is on the difference to the correct value.

## Mean absolute error (MAE)

- The mean difference of predicted values  $y_i$  to ground-truth values  $\hat{y}_i$
- MAE does not treat outliers specifically.

$$MAE := \frac{1}{n} \cdot \sum_{i=1}^n |y_i - \hat{y}_i|$$

## Mean squared error (MSE)

- The mean squared difference of predicted to ground-truth values
- MSE is sensitive to outliers.

Sometimes, the root mean squared error (RMSE) is used:  $RMSE = \sqrt{MSE}$ .

$$MSE := \frac{1}{n} \cdot \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

# Effectiveness Measures for Synthesis Tasks

## Evaluating synthesis tasks

- Goal: Judge quality of generated texts.
- Problem: Multiple outputs may be correct.

**Ground truth.** “Abolish death penalty forever.”  
**Generated.** “Let’s ban the death penalty.”

## Two types of evaluation

- **Automatic.** Quantify similarity between ground truth and generated text.
- **Manual.** Human annotators assess the quality of generated texts.

## Dilemma

- Only manual evaluation is seen as reliable, but it costs time and money.
- Automatic evaluation is needed to observe progress while developing.

# Effectiveness Measures for Synthesis Tasks

## Overview of Measures

### Automatic evaluation measures

- **BLEU**. Precision of  $n$ -gram overlap with brevity penalty
- **ROUGE**. Recall of  $n$ -gram overlap, either for a specific  $n$  or averaged
- **BERTScore**.  $F_1$  derived from similarity matching of text embeddings

Formulas left out here for brevity

### Manual evaluation measures

- Usually, quality dimensions are scored on a Likert scale (say, 1–5).
- The mean or majority judgment of annotators is used for evaluation.

Sometimes, different candidates are also ranked relatively.

### Selected quality dimensions for synthesis tasks

- **Syntax**. Gramaticality, fluency, naturalness, ...
- **Semantics**. Meaning preservation, coherence, ...
- **Pragmatics**. Relevance, informativeness, ...

What dimensions to assess, depends on the task.

# Empirical Experiments

# Empirical Experiments

## Empirical experiments in NLP

- An empirical experiment tests a hypothesis based on observations.
- The focus is here on effectiveness evaluation in NLP.

## Intrinsic vs. extrinsic evaluation

- **Intrinsic.** The effectiveness of an approach is directly evaluated on the task it is made for.

“What accuracy does a part-speech tagger  $XY$  have on the dataset  $D$ ?”

- **Extrinsic.** The effectiveness of an approach is evaluated by measuring how effective its output is in a downstream task.

“Does the output of  $XY$  improve sentiment analysis on  $D$ ?”

## Corpus-based experiments vs. user studies

- We consider the empirical evaluation of approaches on *corpora* here.
- A whole different branch of experiments is related to *user studies*.

Not covered in this course

# Empirical Experiments

## Text Corpora

### Text corpus (plural text *corpora*)

- A principled collection of (mostly real-world) natural language texts with known properties, compiled to study a language problem

Examples: 200,000 product reviews for sentiment analysis,  
1000 news articles for part-of-speech tagging, ...

- The texts in a corpus are often annotated, at least for the problem to be studied.

Examples: Sentiment polarity of a full text,  
part-of-speech tags of each token, ...



<https://pixabay.com>

### Dataset

- A subset of a corpus used for development or evaluation
- NLP methods are trained and tested on the datasets of a corpus.
- Without a corpus, it is hard to develop a strong method — and even harder to reliably evaluate it.

# Empirical Experiments

## Annotations

### Annotation

- An annotation marks a text or a span of text as representing meta-information of a specific type.
- It may also be used to specify relations between other annotations.

**Time entity**                      **Organization entity**  
“ 2014 ad revenues of Google are going to reach  
\$20B. The **Reference** **Time entity**  
search company was founded in '98.  
**Reference**                      **Time entity**                      **Founded relation**  
Its IPO followed in 2004. [...] “

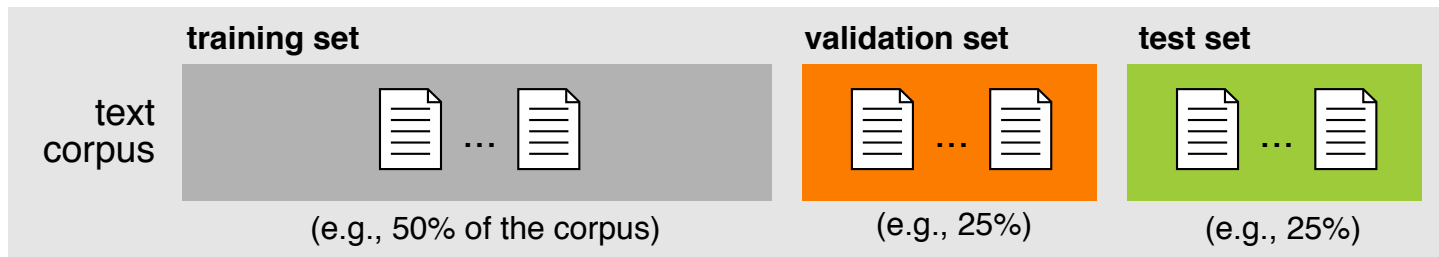
**Topic:** "Google revenues"    **Genre:** "News article"

### Manual vs. automatic annotation

- **Manual.** Most corpora are annotated by human experts or lay persons. NLP methods are developed based on such *ground-truth* annotations.
- **Automatic.** Technically, many NLP methods add annotations to texts.

# Empirical Experiments

## Training, Validation, and Test



### Training set

- Known instances used to develop or statistically learn an approach
- The training set may be analyzed manually and automatically.

### Validation set (aka development set)

- Unknown test instances used to iteratively evaluate an approach
- The approach is optimized on (and adapts to) the validation set.

### Test set (aka held-out set)

- Unknown test instances used for the final evaluation of an approach
- The test set represents unseen data.

# Empirical Experiments

## Cross-Validation



### $n$ -fold cross-validation

- Data is split into  $n$  dataset folds of equal size, often  $n = 10$ .
- The evaluation results are averaged over  $n$  runs.

### Training and evaluation

- In the  $i$ -th run, the  $i$ -th fold is used for evaluation (validation).
- All other folds are used for development (training).

### Cross-validation + test set

- When doing cross-validation, a held-out test set is still important.
- Otherwise, repeated development will overfit to the splitting.

# Empirical Experiments

## Comparison

### Why comparing?

- A new method is seen as useful, if it is better than other methods, usually measured in terms of effectiveness.
- To test this, methods are compared to *baselines*.

### Baseline

- A baseline is an alternative method that has been proposed before or that can easily be realized.
- Ideally, a new method should be better than all baselines.

### Types of baselines

- **Trivial.** Methods that can easily be derived from a given task or dataset
- **Standard.** Methods that are often used for related tasks
- **Ablation.** Sub-methods of a newly proposed method
- **State of the art.** The best published method for the task (if available).

# Empirical Experiments

## Descriptive Statistics

### Descriptive statistics

- Methods for summarizing a sample  $\tilde{X}$  (or distribution  $X$ ) of values in order to *describe phenomena*

### Selected measures

- **Mean.** The arithmetic average  $M$  of a sample of values  $\tilde{X}$  of size  $n$ .  $M$  is used for a sample,  $\mu$  for the whole distribution.

$$M := \frac{1}{n} \sum_{i=1}^n \tilde{X}_i$$

- **Variance.** The mean  $s^2$  of all values' squared differences to the mean.  $s$  is used for a sample,  $\sigma$  for the whole distribution.

$$s^2 := \frac{1}{n-1} \sum_{i=1}^n (\tilde{X}_i - M)^2 \quad \sigma^2 := \frac{1}{n} \sum_{i=1}^n (\tilde{X}_i - \mu)^2$$

- **Standard deviation.** The square root of the variance

$$s = \sqrt{s^2} \quad \sigma = \sqrt{\sigma^2}$$

# Empirical Experiments

## Inferential Statistics

### Inferential statistics

- Methods for drawing conclusions based on values in order to *generalize inferences* beyond a given sample  $\tilde{X}$

### Two competing hypotheses

- **Research hypothesis ( $H$ )**. Prediction about how a change in variables will cause changes in other variables

“There is **a statistically significant difference** between the RMSE of our approach and the RMSE reported by Persing et al. (2015).”

- **Null hypothesis ( $H_0$ )**. Antithesis to  $H$

“There is **no statistically significant difference** between the RMSE of our approach and the RMSE reported by Persing et al. (2015).”

- If  $H_0$  is true, then any results observed in an experiment that support  $H$  are due to chance or sampling error.

# Hypothesis Testing

## Significance test (aka hypothesis test)

- A statistical procedure that determines how likely it is that the results of an experiment are due to chance (or sampling error)
- It tests whether a null hypothesis  $H_0$  can be rejected (and hence,  $H$  can be accepted) at some chosen *significance level*.

## Significance level $\alpha$

- The accepted risk (in terms of a probability) that  $H_0$  is wrongly rejected
- A choice of  $\alpha = 0.05$  means that there is at least a 95% chance that a potential rejection of  $H_0$  is correct.

Usually,  $\alpha$  is set to 0.05 (default) or to 0.01.

## *p*-value

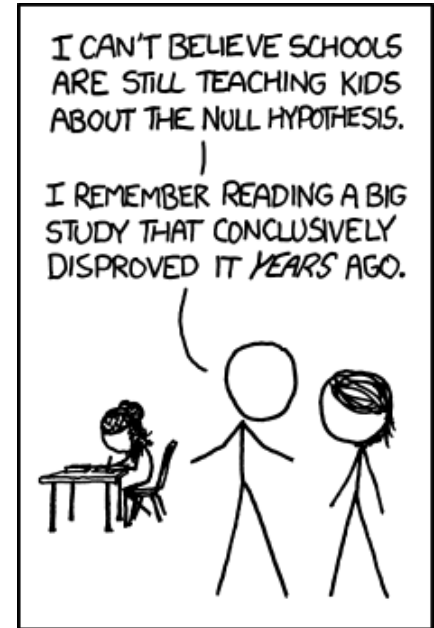
- The likelihood (in terms of a probability) that results are due to chance.
- If  $p \leq \alpha$ ,  $H_0$  is rejected. The results are seen as statistically significant.
- If  $p > \alpha$ ,  $H_0$  cannot be rejected.

# Hypothesis Testing

## Carrying out a Significance Test

### Main test steps

1. **Hypothesis.** State  $H$  and  $H_0$ .
2. **Significance level.** Choose  $\alpha$  (*before* the test).
3. **Testing.** Carry out a significance test, which fits the data, to get the  $p$ -value.
4. **Decision.** Reject  $H_0$  or fail to reject it.



### Significance tests

- Different tests exist that make different assumptions about the data.
- **Parametric.** More likely to detect a significant effect when one exists
- **Non-parametric.** Fewer assumptions and, thus, more often applicable

<b>Parametric test</b>	<b>Non-parametric correspondent</b>
Independent student's $t$ -test	Mann-Whitney Test
Dependent and one-sample student's $t$ -test	Wilcoxon Signed-Rank Test
...	...

# Hypothesis Testing

## Assumptions of Significance Tests

### Assumptions of all tests

- **Sampling.** The sample is a random sample from the distribution.  
Notice: In NLP, each “instance” of a sample usually consists of multiple texts.
- **Values.** The values within each variable are independent.

### Assumption of all parametric tests

- **Scale.** The dependent variable has an interval or ratio scale.
- **Distribution.** The given variables are normally distributed.
- **Variance.** Distributions that are compared have the same variances.

### Test-specific assumptions

- In addition, specific tests may have specific assumptions.

### Example: Student's $t$ -test

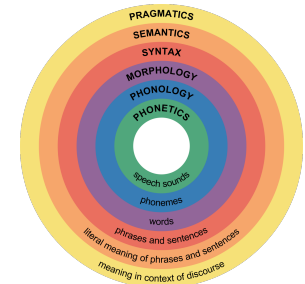
- The independent variable has a nominal scale.
- $t$ -tests are robust over moderate violations of the normality assumption.

# Conclusion

# Conclusion

## NLP and linguistics

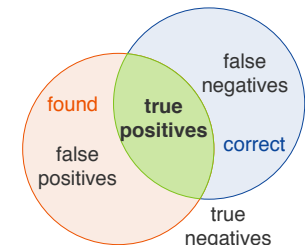
- Linguistic knowledge from phonetics to pragmatics
- Ambiguity exists across linguistic levels
- Techniques from simple rules to language models



<https://en.wikipedia.org>

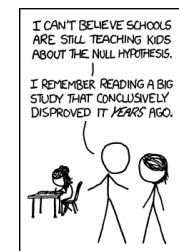
## Measures

- Accuracy, precision, recall, and  $F_1$ -score for classes
- Mean absolute error and mean squared error for values
- Measures like BLEU or human judgment for text



## Experiments

- Methods are developed evaluated on corpus datasets
- Experiments compare methods against baselines
- Statistical tests explore and “prove” quality of methods



<https://pixabay.com>  
<https://xkcd.com/892/>

# References

## Some content taken from

- **Jurafsky and Manning (2016)**. Daniel Jurafsky and Christopher D. Manning. Natural Language Processing. Lecture slides from the Stanford Coursera course, 2016.  
<https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html>
- **Rockinson-Szapkiw (2013)**. Amanda J. Rockinson-Szapkiw. Statistics Guide, 2013.  
<http://amandaszapkiw.com/elearning/statistics-guide/downloads/Statistics-Guide.pdf>
- **Wachsmuth (2015)**. Henning Wachsmuth. Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer, 2015.
- **Wachsmuth (2024)**. Henning Wachsmuth. Introduction to Natural Language Processing. Lecture slides, 2024.  
<https://www.ai.uni-hannover.de/en/teaching/courses/inlp>