Language Processing

Einführung in die Informationsverarbeitung

Nils Reiter

December 1, 2022

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Today

Computational Linguistics

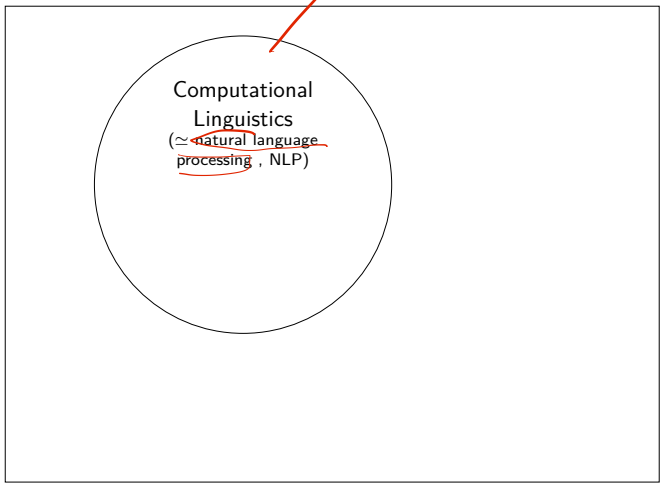
- Experiments
- Manual Annotation
- Automatization

Summary

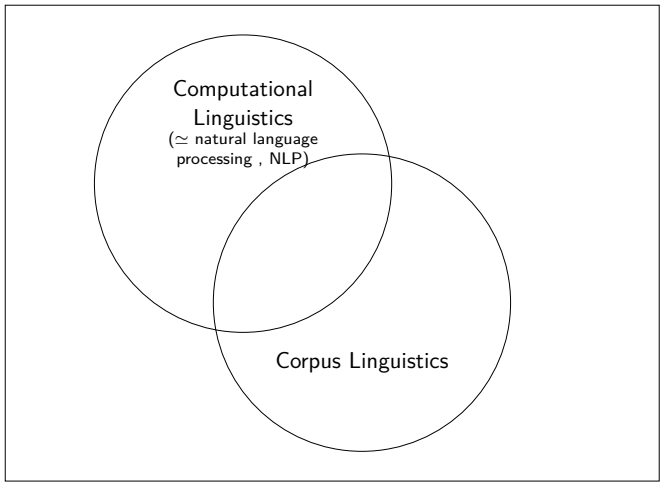
Section 1

Computational Linguistics

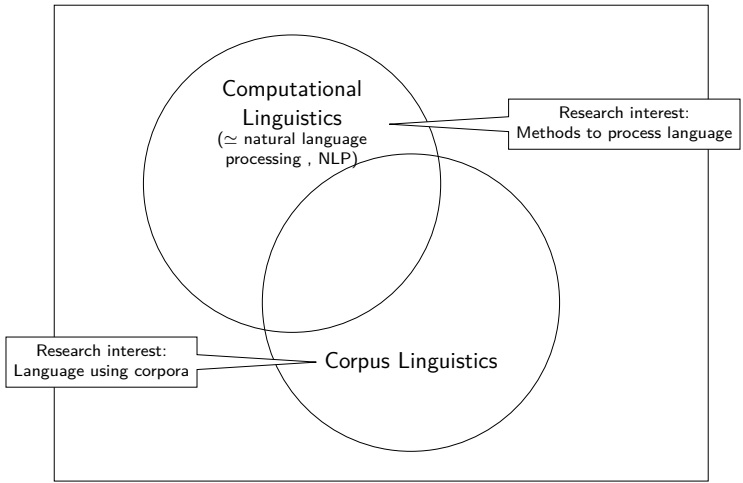
Disciplinary Placement



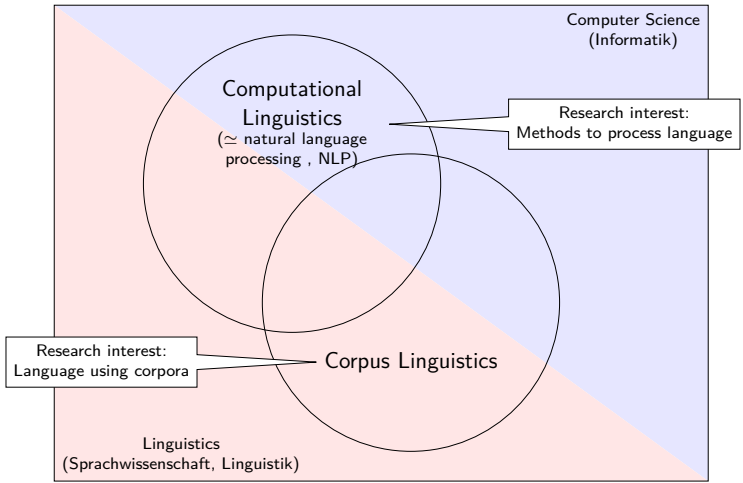
Disciplinary Placement



Disciplinary Placement



Disciplinary Placement



Brief history of Computational Linguistics I

- 1950s: DARPA Projects to automatically translate Russian into English
- 1957/65: Linguistics shifts focus from describing to generating Chomsky (1957, 1965)
- 1959: Theo Lutz for the first time generates a German poem with a computer Lutz (1959)
- 1962: Foundation of the »Association for Machine Translation and Computational Linguistics«, 1968 renamed to »Association for Computational Linguistics (ACL)«
- 1966, ALPAC report: MT more expensive, less accurate and slower than human translation ALPAC (1966)
- 1968: Foundation of SYSTRAN, first MT company
- 1975: European commission uses SYSTRAN software (first use of MT on EU level)

Brief history of Computational Linguistics II

- 1984: First corpus-based commercial MT system Nagao (1984)
- 1992: Study programs established in Germany (Saarbrücken/Stuttgart)
- 2011: IBM Watson beats two humans in Jeopardy: `https://www.youtube.com/watch?v=WFR3lOm_xhE` / Apples Siri launched
- 2013: Word embeddings (e.g., word2vec) Mikolov et al. (2013)
- 2017: Launch of the DeepL Translator
- 2018: Transformer models: BERT Devlin et al. (2019)

Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ›digital methods‹ on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, p. 4)
  - … but this is a minority position in CL, often also seen as part of AI

Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ›digital methods‹ on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, p. 4)
  - … but this is a minority position in CL, often also seen as part of AI
  - Historically (and still today) split between engineering (natural language processing, NLP) and science/scholarship (computational linguistics, CL)
  - ⚠ Neurolinguistic programming and natural language processing are not the same (both use ›NLP‹ as abbreviation)

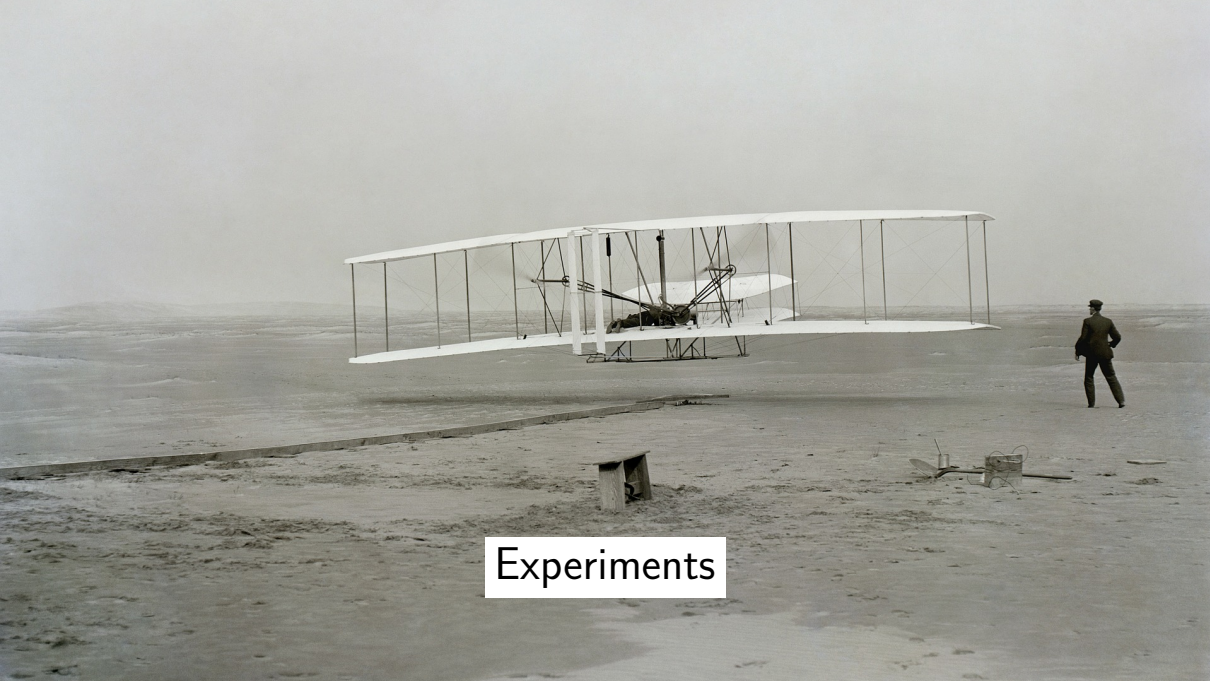
Digital Humanities and Computational Linguistics

Today

- Digital Humanities, broadly: Working with ›digital methods‹ on humanities subjects
- Linguistics: Study of language
- Computational Linguistics: Pioneer DH area Reiter (2014, p. 4)
  - … but this is a minority position in CL, often also seen as part of AI
  - Historically (and still today) split between engineering (natural language processing, NLP) and science/scholarship (computational linguistics, CL)
  - ⚠ Neurolinguistic programming and natural language processing are not the same (both use ›NLP‹ as abbreviation)

University of Cologne

For historic reasons, CL and NLP are called »Sprachliche Informationsverarbeitung«


Experiments

Experiments

- Cornerstone of the ›scientific method‹
- Used in many disciplines: Natural sciences, social sciences, medicine, ...

Experiments

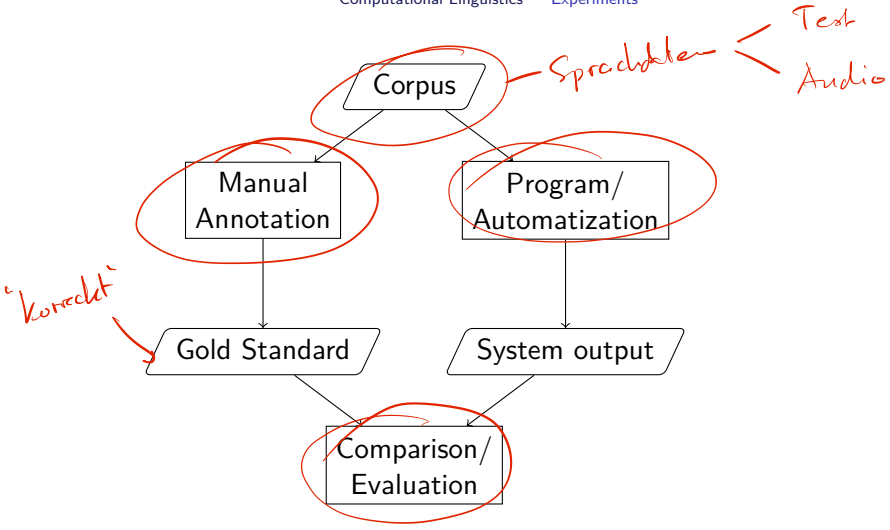
- Cornerstone of the ›scientific method‹
- Used in many disciplines: Natural sciences, social sciences, medicine, ...
- Experiments are used to verify or falsify hypotheses
- Reproducibility: The outcome does not depend on the experimenter

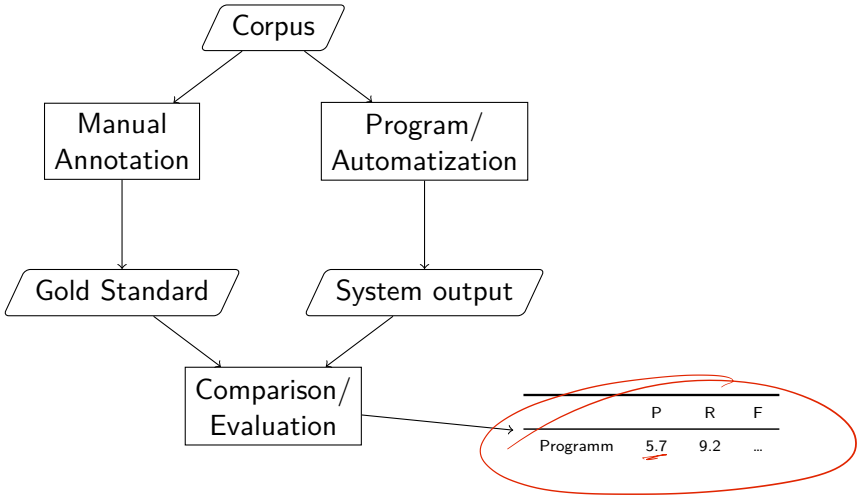
Experiments

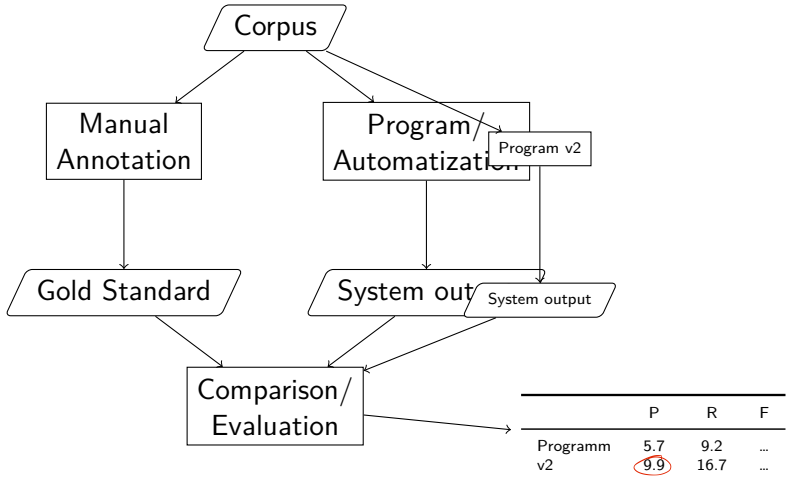
- Cornerstone of the ›scientific method‹
- Used in many disciplines: Natural sciences, social sciences, medicine, ...
- Experiments are used to verify or falsify hypotheses
- Reproducibility: The outcome does not depend on the experimenter
- CL: Hypotheses about the operationalisation of language/text phenomena

Example

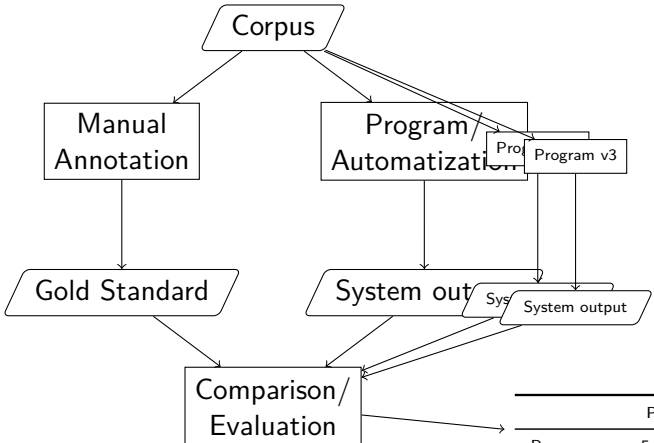
Position within a sentence is indicative for the part of speech


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation

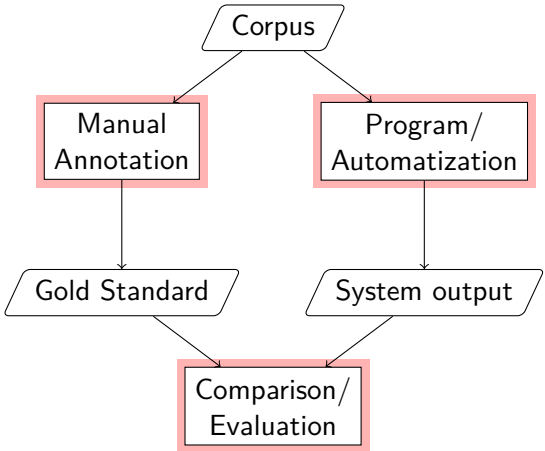

Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation
P
R
F
Programm
5.7
9.2
...

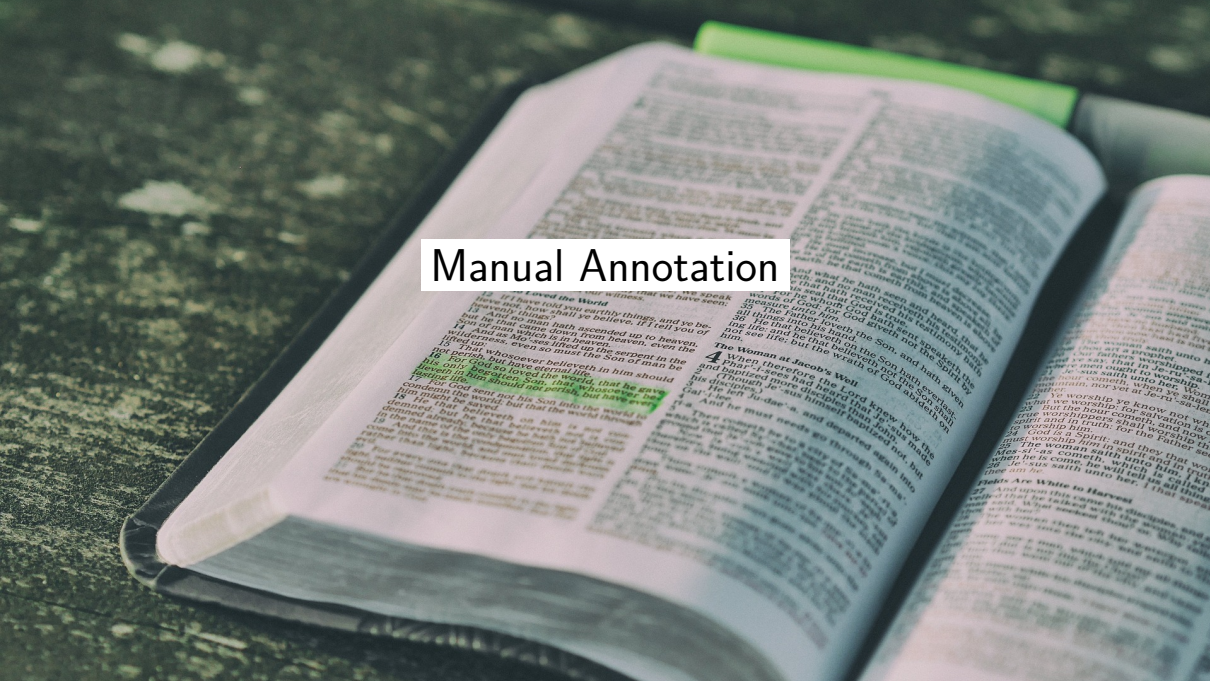


	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...



	P	R	F
Programm	5.7	9.2	...
v2	9.9	16.7	...
v3	15.3	21.8	...


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation


Manual Annotation

Annotation

- Interdisciplinary ›false friend‹
- Different meanings in different disciplines
  - Adding TEI/XML markup: DH community
  - Adding comments to page margins: Hermeneutic traditions
    - Literary studies, bible studies
  - Assigning categories to textual material: (computational) linguistics

Annotation

- Interdisciplinary ›false friend‹
- Different meanings in different disciplines
  - Adding TEI/XML markup: DH community
  - Adding comments to page margins: Hermeneutic traditions
    - Literary studies, bible studies
  - Assigning categories to textual material: (computational) linguistics

Example

Der	alte	Mann	wird	verrückt	.

Annotation

- Interdisciplinary ›false friend‹
- Different meanings in different disciplines
  - Adding TEI/XML markup: DH community
  - Adding comments to page margins: Hermeneutic traditions
    - Literary studies, bible studies
  - Assigning categories to textual material: (computational) linguistics

Example

Der	alte	Mann	wird	verrückt	.
Artikel	Adjektiv	Nomen	Verb	?	Punkt

Annotation Guidelines

/ "Codebuch"

- Describe the way to create the machine-readable truth
- What is to be annotated (which words)
- Working definitions or tests for categories
- Living documents: Need to be iteratively improved
- Community-wide accepted standards are needed

Stuttgart-Tübingen Tagset (STTS)

Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS

Anne Schiller, Simone Teufel, Christine Stöckert
Universität Stuttgart
Institut für maschinelle Sprachverarbeitung

Christine Thielen
Universität Tübingen
Seminar für Sprachwissenschaft

Draft

14. November 1995

1.1 Zuweisung von Tags

Als allgemeine Regel gilt, daß jede Wortform genau ein Tag enthält. Der Begriff Wortform umfaßt neben "echten" Wortformen auch Zahlen in Ziffern, Satzzeichen, Sonderzeichen (wie z.B. §, $), abgetrennte Wortteile oder Kompositions-Erstglieder (wie z.B. **Ein-** und Ausgang) etc. Es wird davon ausgegangen, daß für das manuelle Taggen die Texte so aufbereitet sind, daß jede Zeile genau eine Wortform enthält.

1.2 Mehrwortlexeme

Damit ist es also (aus technischen Gründen) nicht möglich, Mehrwortlexeme als Ganzes zu taggen, oder kontraktive Formen mit einer Kombination aus mehreren Tags zu versehen. Idealerweise sollten feststehende Ausdrücke wie *vor kurzem*, *vor allem* als Mehrwortlexeme (**multi word items**) aufgefaßt werden und von Tokenizer und Tagger so behandelt werden. Solange dies technisch noch nicht möglich ist, werden als Kompromiß die einzelnen Teile annähernd so behandelt, als wenn die Teile einzeln stehen würden:

Beispiele:

- New/NE **York**/NE — **nichts** New **York**/NE
- so/ADV **etwas**/KOUS — **nichts** so **etwas**/KOUS
- zum/APPRART — **nichts** zum/APPR ART

Bei aus 2 Teilen bestehenden Konjunktionen (*entweder – oder*, *weder – noch*) werden **beide** Teile als KON getaggt. In den folgenden guidelines werden Mehrwortlexeme durch das Zeichen **=**| gekennzeichnet, was besagt, daß diese Wortform idealerweise ein gemeinsames Tag bekommen sollte (welches hinter dem Zeichen **=**| angegeben wird), als Kompromißlösung aber wie angegeben getaggt wird.

1.3 Behandlung von Abkürzungen

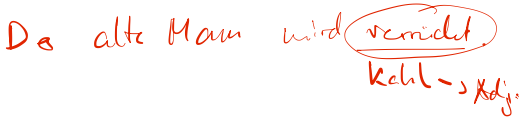
Es gibt kein eigenes Tag für Abkürzungen. Abgekürzte Wortformen werden generell so getaggt wie die ausgeschriebene Form. Abkürzungen für mehrere Worte, die nicht durch Leerzeichen getrennt sind, werden entsprechend ihrer syntaktischen Funktion klassifiziert.

Beispiele:

- Herr/NN **Dr.**/NN **Meier**/NE
- *die* gem./ADJA *Verhandlungen*
- *mit Haus* u./KON *Garten*
- z./APPRART **Z.**/NN
- z.B./ADV
- d./PDS h./WFIN
- d.h./KON
- sondern/KON

Stuttgart-Tübingen Tagset (STTS)

- Welche Wortart hat das Wort ›verrückt‹?



Stuttgart-Tübingen Tagset (STTS)

- Welche Wortart hat das Wort ›verrückt‹?

Verb weil es Partizip von ›verrücken‹ ist

Adjektiv weil es gesteigert werden kann: ›verrückt – verrückter – am verrücktesten‹

Stuttgart-Tübingen Tagset (STTS)

- Welche Wortart hat das Wort ›verrückt‹?

  Verb weil es Partizip von ›verrücken‹ ist

  Adjektiv weil es gesteigert werden kann: ›verrückt – verrückter – am verrücktesten‹

- Es ist beides: ›verrückt‹ ist ambig (bzgl. seiner Wortart)
  - »Ich habe verrückt.« vs. »Ich werde verrückt.«

Stuttgart-Tübingen Tagset (STTS)

- Welche Wortart hat das Wort ›verrückt‹?

  Verb weil es Partizip von ›verrücken‹ ist

  Adjektiv weil es gesteigert werden kann: ›verrückt – verrückter – am verrücktesten‹

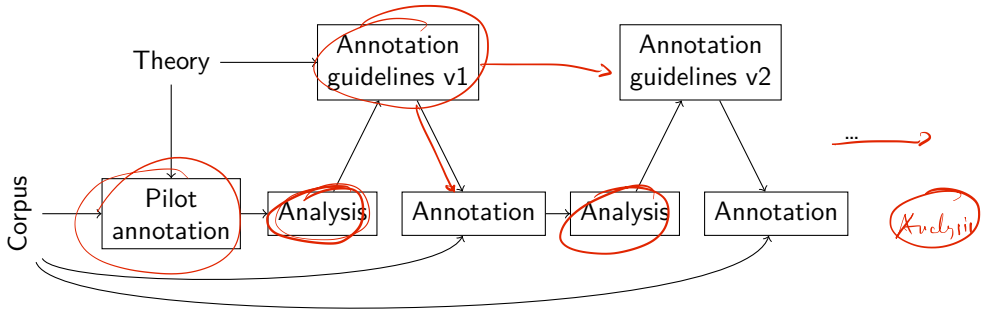
- Es ist beides: ›verrückt‹ ist ambig (bzgl. seiner Wortart)
  - »Ich habe verrückt.« vs. »Ich werde verrückt.«

Kriterien für Disambiguierung Kopulakonstruktionen mit ADJD vs. Verlaufspassiv mit VVPP:

- Verdacht auf VVPP: kann der Satz ins Aktiv gesetzt werden mit gleicher Semantik? Ja → VVPP
- von–PP oder ähnliche PP, die auf Verbsemantik hinweist → VVPP
- Ersetzung durch semantisch nahes Adjektiv möglich → ADJD

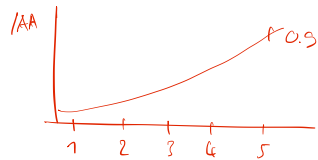
Abbildung: STTS Guidelines, S. 23

Annotation Workflow



Hovy and Lavid (2010); Reiter (2020)

Annotation Analysis



- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified (›Inter-Annotator-Agreement‹, IAA)

Cohen 1960; Fleiss 1971; Fournier 2013; Mathet et al. 2015

- Inter- und Intra-AA
- ...It's also a good idea to talk to the annotators

Indirect Annotations

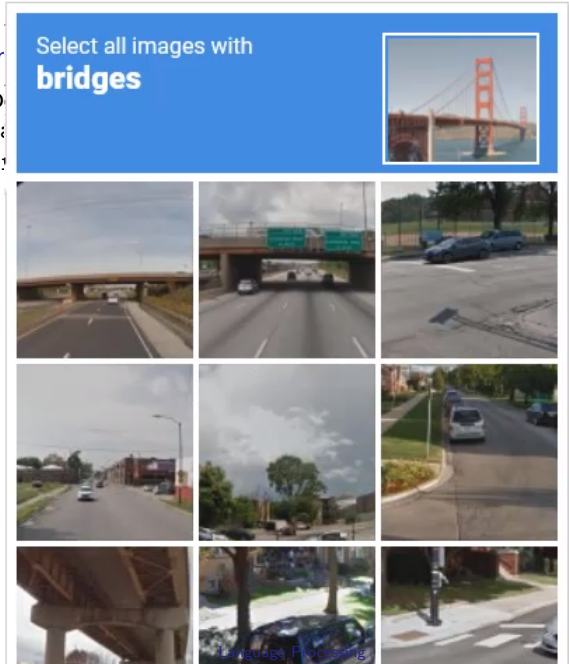
- Annotations as a by-product of games
  - https://www.artigo.org
  - https://anawiki.essex.ac.uk/phrasedetectives/

Indirect Annotations

- Annotations as a by-product of games
  - `https://www.artigo.org`
  - `https://anawiki.essex.ac.uk/phrasedetectives/`
- Captchas for OCR correction

Indirect Annotation

- Annotations as a b
  - https://www.a
  - https://anaw
- Captchas for OCR



Standard Corpora

- Iterative improvements of programs
- Measuring these improvements

- Accessibility
- Acceptance in the research community

Learning from Raw Data

Corpus annotation is expensive, therefore:

Learning from Raw Data

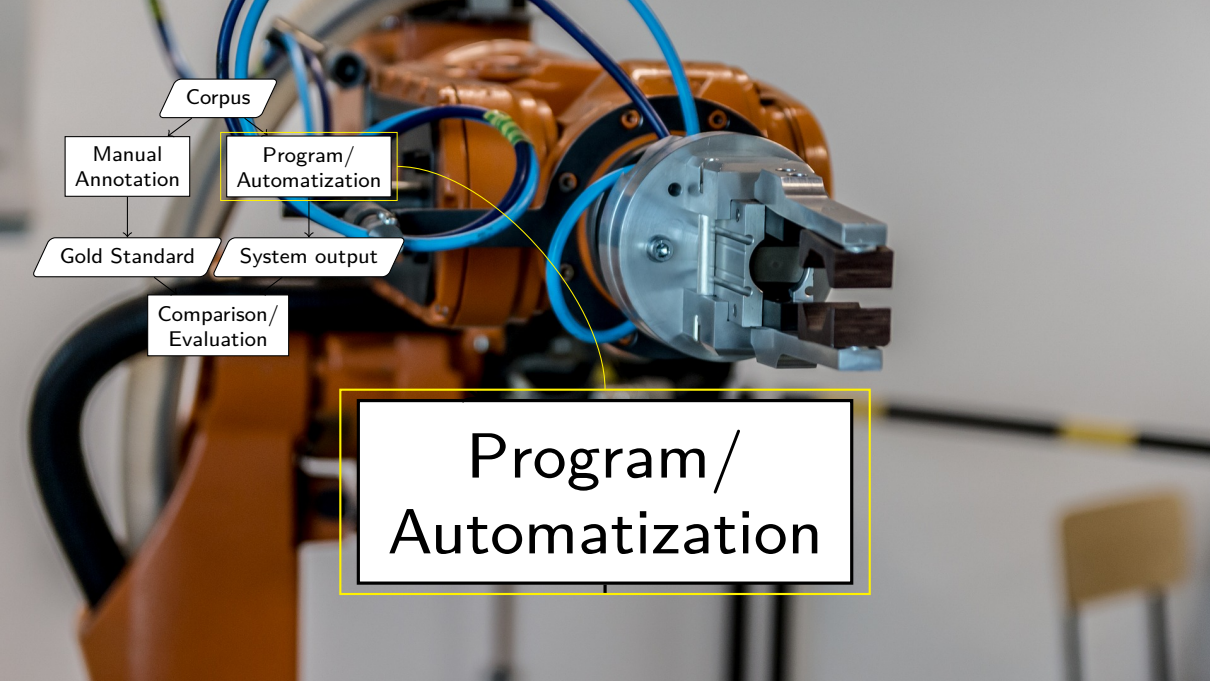
Corpus annotation is expensive, therefore:

- Train on things that are already there
- word2vec: Is ›dog‹ a context word of ›lazy‹? Mikolov et al. (2013)
- BERT Devlin et al. (2019)
  - Can you fill in this blanked word? (»masked language modeling«, MLM)
  - Are these two sentences natural neighbours? (»next sentence prediction«, NSP)

Learning from Raw Data

Corpus annotation is expensive, therefore:

- Train on things that are already there
- word2vec: Is ›dog‹ a context word of ›lazy‹? Mikolov et al. (2013)
- BERT Devlin et al. (2019)
  - Can you fill in this blanked word? (»masked language modeling«, MLM)
  - Are these two sentences natural neighbours? (»next sentence prediction«, NSP)
- Training data available in abundance
  - As long as there is digital data for a language
  - ⚠ Difficult to control what exactly is in there
    - More obvious for text-image data sets Birhane et al. (2021) haveibeentrained.com


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation
Program/ Automatization

Systems

= Programs, models, ...

- Predict annotations
- Idealy: The same annotations as a human (the ›correct‹ ones)

Systems

= Programs, models, ...

- Predict annotations
- Idealy: The same annotations as a human (the ›correct‹ ones)

Types

- Rule-based (popular 1950s – 2010s)
- Statistical (popular 2005 – 2015)
  - Now often called ›classical machine learning‹
- Neural (popular since 2015)
  - Also often called ›deep learning‹

Rule-based Systems

- Manually specified rules over certain criteria
  - e.g., HPSG grammar
- Criteria: Vocabulary from which rules are created
  - e.g., Noun: Every token, that starts with an upper case letter
  - e.g., Noun: Every token, that starts with an upper case letter and is not sentence initial

toCharArray()[0] => char

Supervised Systems

"Classical ML"

- Classification: Associate items with previously known categories
- Learn patterns from annotated data (= training data)
- Relations between input (represented as a vector of feature values, X) and output (Y)
  - Can be an n-to-m relation, but mostly n-to-1 (i.e., we predict a single target category)

(Casing, Länge, -en) => Wortart

Der	3	Groß	0	Artikel
Hund	4	Klein	0	Nomen

Features

- The properties of a item that is to be classified
- Classical machine learning
  - Manual coding of explicit, scientifically validated features: Feature extraction
  - »Translation« of the corpus into feature vectors
  - Feature engineering
  - Design and implementation of feature extractors
  - Linguistic features need to be determined somehow
    → Dependencies, modularization

Features

- The properties of a item that is to be classified
- Classical machine learning
  - Manual coding of explicit, scientifically validated features: Feature extraction
  - »Translation« of the corpus into feature vectors
  - Feature engineering
  - Design and implementation of feature extractors
  - Linguistic features need to be determined somehow
    $\rightarrow$ Dependencies, modularization
- Deep learning
  - Embeddings used as features
    - A word is mapped onto an n-dimensional vector, which is then put into the ML system
    - Vector dimensions = features
    - But not so helpful anymore

Example: Parts of Speech

Features	Data type
Case	Binary
Length	> 0

Table: Features

Token	Case	L.
Der	u	3
Hund	u	4
bellt	l	5
.	?	1
Die	u	3
Katze	u	5
schnurrt	l	8
.	?	1

Table: Feature extraction

Example: Parts of Speech

Feature	Data type
Case	Binary
Length	> 0
Sentence initial	Binary

Table: Features

Token	Case	L.	S. initial
Der	u	3	Y
Hund	u	4	N
bellt	l	5	N
.	Jein	?	N
Die	u	3	Y
Katze	u	5	N
schnurrt	l	8	N
.	?	1	N

Table: Feature extraction

Example: Parts of Speech

Feature	Data type
Case	Binary
Length	> 0
Sentence initial	Binary

Tabl… Features

Introduces dependency!

Token	Case	L.	S. initial
Der	u	3	Y
Hund	u	4	N
bellt	l	5	N
.	Jein	?	N
Die	u	3	Y
Katze	u	5	N
schnurrt	l	8	N
.	?	1	N

Table: Feature extraction

Embeddings

- Classical machine learning
  - Instance is represented by a feature vector
  - Features are humanly interpretable properties of the instance

Embeddings

- Classical machine learning
  - Instance is represented by a feature vector
  - Features are humanly interpretable properties of the instance
- Embeddings
  - Instance is represented by a vector in an n-dimensional space
  - Mapping of instance to vector is learned, i.e., part of the training process

Embeddings

- Classical machine learning
  - Instance is represented by a feature vector
  - Features are humanly interpretable properties of the instance
- Embeddings
  - Instance is represented by a vector in an n-dimensional space
  - Mapping of instance to vector is learned, i.e., part of the training process

Example (Vector for »köln«)

0.0539 -0.0030 0.0203 -0.1084 -0.0099 0.0705 -0.0546 -0.0433 -0.0096 0.0561 -0.0095 0.0280 0.1726 0.0190 0.0369 0.0217 -0.0002 -0.0309 0.0347 -0.0749
-0.0202 0.0151 -0.0195 0.0001 0.0232 0.0243 -0.0170 -0.0090 -0.0108 -0.0943 0.0376 0.1118 -0.0324 0.0148 -0.0033 0.0537 -0.0681 -0.0733 -0.0201 -0.0329
0.1242 0.0324 -0.0744 -0.0149 -0.0047 -0.0484 -0.0483 0.0481 0.0107 0.0101 -0.0704 0.0500 0.0112 -0.0227 0.0499 -0.0259 -0.0441 0.0712 -0.0157 -0.1271
0.0407 -0.0495 -0.0359 0.0202 0.0024 0.0764 0.0196 0.0267 -0.0117 0.0026 0.0171 -0.0121 -0.1374 -0.0370 0.0247 -0.0113 -0.0094 0.0322 -0.0347 -0.0866 0.0042
-0.0014 0.0067 0.0591 0.0009 0.0085 0.0310 0.0479 -0.0511 0.0198 -0.0886 -0.0274 -0.1364 0.0322 -0.1638 -0.0689 0.0016 -0.1039 0.0059 0.0757 -0.0034 0.1013
-0.0034 -0.0065 -0.0468 0.1577 -0.0065 -0.0478 -0.0004 0.0682 0.0045 -0.0607 -0.0590 0.0343 0.0036 -0.1014 -0.0136 -0.0063 0.0801 0.0360 0.0579 -0.0039
0.0975 0.0500 -0.0558 -0.0095 0.0057 -0.0246 0.1070 -0.0186 0.0669 -0.0781 -0.0569 -0.1286 -0.0834 0.0106 -0.0672 -0.0205 0.0613 0.0290 -0.0545 -0.0481
-0.0882 -0.0489 0.0622 -0.0730 -0.0192 -0.0415 -0.0287 0.0218 -0.0427 -0.0046 0.0255 -0.1164 0.0077 -0.0546 -0.0786 0.0000 -0.0456 0.0943 0.0157 -0.0117
-0.0441 -0.0015 -0.0556 -0.0508 0.0088 0.0418 0.0030 -0.1450 -0.0663 0.0800 0.0172 -0.0289 0.1178 -0.0973 0.0888 0.0637 -0.0295 0.0212 0.0100 -0.0860 0.0035
0.0730 0.0425 -0.0080 0.0885 -0.0166 -0.0765 0.0004 -0.0118 0.0138 -0.0093 -0.0606 -0.0447 -0.0746 0.0131 -0.0447 -0.0763 0.0032 0.1181 0.0542 0.0431
-0.0273 0.0547 0.0135 0.0006 -0.0241 -0.0418 0.0278 -0.0821 -0.0572 -0.0039 0.0214 -0.0196 0.0449 -0.0286 0.0204 0.0681 -0.0901 -0.0266 -0.0287 -0.0874
0.0797 -0.0784 -0.0920 0.0380 0.0411 0.0859 0.0369 0.0595 0.0446 0.0363 -0.0353 -0.0044 -0.0061 0.1134 0.1420 -0.0026 -0.0013 0.0033 0.0508 0.0096 -0.0757
0.0085 -0.0099 -0.0384 0.0218 -0.0259 -0.0112 -0.0212 0.0273 0.0532 -0.0278 -0.0634 0.0317 -0.0022 0.0882 -0.0240 0.0031 -0.0370 0.0747 -0.0097 -0.0315
0.0405 0.0124 -0.1416 -0.0768 0.0363 -0.1248 -0.0134 0.0702 -0.0905 -0.0387 0.0683 -0.0784 0.0886 0.0640 0.0611 -0.0199 -0.0447 -0.1331 -0.1247 0.0540
0.0499 -0.0212 -0.0544 -0.1161 -0.0729 0.0894 0.0532 0.0164 -0.0039 -0.0108 -0.0248 -0.1021 -0.0549 -0.0318 0.0309 -0.0691


Summary

Summary

- Experimental science
  - Real-world applications, from smart assistants over machine translation to text generation
- Annotations as coded, machine-readable ›nature‹
  - Annotation decisions based on annotation guidelines
  - Inter-annotator agreement for validation
- Programs to automatically detect/categorize linguistic phenomena
  - Rule-based: Manually crafted rules between input and output
  - Classical machine learning: Computer learns rules based on manually crafted features
  - Deep learning: Computer also learns its own input representation

References I

- ALPAC. *Language and Machines. Computers in Translation and Linguistics*. Tech. rep. National Research Council, 1966.
- Austin, John Langshaw. *How to Do Things with Words*. William James lectures. Harvard University Press, 1962.
- Bar-Hillel, Yehoshua. »Out of the pragmatic wastebasket«. In: *Linguistic Inquiry* 2 (1971), pp. 401–407.
- Bernhart, Toni. »Beiwerk als Werk. Stochastische Texte von Theo Lutz«. In: *editio* 34 (2020).
- Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. »Multimodal datasets: misogyny, pornography, and malignant stereotypes«. In: *CoRR* abs/2110.01963 (2021). arXiv: 2110.01963. URL: https://arxiv.org/abs/2110.01963.
- Chomsky, Noam. *Syntactic Structures*. Mouton De Gruyter, 1957.
- — .*Aspects of the theory of syntax*. MIT Press, 1965.

References II

- Cohen, Jacob. »A Coefficient of Agreement for Nominal Scales«. In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.
- Cortes, Corinna and Vladimir Vapnik. »Support-vector networks«. In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297.
- Davidson, Donald. »Truth and meaning«. In: *Synthese* 17.1 (1967), pp. 304–323.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. »BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding«. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.
- Echelmeyer, Nora, Nils Reiter, and Sarah Schulz. »Ein PoS–Tagger für „das" Mittelhochdeutsche«. In: *Book of Abstracts of DHd 2017*. Bern, Switzerland, Feb. 2017. URL: https://elib.uni-stuttgart.de/handle/11682/9040.

References III

- Fleiss, Joseph L. »Measuring nominal scale agreement among many raters«. In: *Psychological Bulletin* 76.5 (1971), pp. 420–428.
- Fournier, Chris. »Evaluating Text Segmentation using Boundary Edit Distance«. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, pp. 1702–1712. URL: http://aclweb.org/anthology/P13-1167.
- Grice, Herbert Paul. »Logic and conversation«. In: *Syntax and Semantics* 3.S 41 (1975). Ed. by P. Cole and J. Morgan, p. 58.
- Hovy, Eduard and Julia Lavid. »Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics«. In: *International Journal of Translation Studies* 22.1 (Jan. 2010).
- Levinson, Stephen C. *Pragmatics*. Cambridge University Press, 1983.
- Lutz, Theo. »Stochastische Texte«. In: *augenblick* 4 (1959), pp. 3–9. URL: https://www.netzliteratur.net/lutz%5C_schule.htm.

References IV

- Manning, Christopher D. and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press, 1999.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. »The Unified and Holistic Method Gamma ( ) for Inter-Annotator Agreement Measure and Alignment«. In: *Computational Linguistics* 41.3 (2015), pp. 437–479.
- Mehl, Matthias R., Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, and James W. Pennebaker. »Are Women Really More Talkative Than Men?« In: *Science* 317 (July 2007).
- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean. »Efficient Estimation of Word Representations in Vector Space«. In: *arXiv cs.CL* (Jan. 2013). URL: https://arxiv.org/pdf/1301.3781.pdf.
- Nagao, Makoto. »A Framework of a Mechanical Translation between Japanese and English by Analogy Principle«. In: *Proc. of the International NATO Symposium on Artificial and Human Intelligence*. Lyon, France: Elsevier North-Holland, Inc., 1984, pp. 173–180.

References V

- *Programmgesteuerte Elektronische Rechenanlage Zuse Z 22 und Z 22/R. Programmierungsanleitung*. Zuse KG. Bad Hersfeld, Western Germany, July 1960.
- Quinlan, J. R. »Induction of Decision Trees«. In: *Machine Learning* 1.1 (Mar. 1986), pp. 81–106.
- Reiter, Nils. »Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms«. PhD thesis. Heidelberg University, Germany, June 2014.
- — .»Anleitung zur Erstellung von Annotationsrichtlinien«. In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter, Axel Pichler, and Jonas Kuhn. Berlin: De Gruyter, July 2020, pp. 193–202.
- Schmid, Helmut. »Probabilistic part-of-speech tagging using decision trees«. In: *Proceedings of the conference on New Methods in Language Processing* 12 (1994).
- Shannon, Claude E. »A mathematical theory of communication«. In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423.
- *Z22/R*. Zuse KG. Bad Hersfeld, Western Germany, Jan. 1960.