

Computerlinguistik I

Vorlesung im WiSe 2018/19
(M-GSW-09)

Prof. Dr. Udo Hahn

Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

<http://www.julielab.de>

Allgemeine Hinweise

- Vorlesung: Do, 10-12h (Fürstengrb.1, SR 275)
- Übung zV: Mo, 8-10h (Fürstengrb.1, SR 275)
 - beginnt am **22.10.**
- Vorlesungsmaterialien im Netz
 - <http://www.julielab.de/> ⇒ „Students“
- **M-GSW-09 besteht aus VL+ÜB und Seminar!**
- Sprechstunde: Mi, 12-13h, bA (FG 30, 004)
- Email: udo.hahn@uni-jena.de
- URL: <http://www.julielab.de>
- Fachliteratur ist überwiegend in Englisch

Bitte ...

- ... Handys/Smartphones ausschalten
- ... 90 Minuten ohne Mail-Check sind möglich
„Digital detox“
- ... kein Picknick



Institut für Germanistische Sprachwissenschaft der FSU Jena

- **Lehrstuhl für Theoretische Linguistik – Grammatiktheorie**
 - Prof. Dr. Peter Gallmann – n.n
- **Lehrstuhl für Angewandte Linguistik – Computerlinguistik**
 - Prof. Dr. Udo Hahn
- **Professur für Pragmatik**
 - Prof. Dr. Pia Bergmann
- **Professur für Phonetik & Sprechwissenschaft**
 - Prof. Dr. Adrian Simpson
- **Professur für Geschichte der deutschen Sprache**
 - Prof. Dr. Eckhard Meineke

Computerlinguistik in Jena (1/2)

- **Institutionell: Teil der Germanistischen Sprachwissenschaft**
 - aber einzelsprachübergreifende Methodik
 - besondere Anwendungsdomänen:
 - Naturwissenschaften: Biologie + Medizin
 - Sozial- und Wirtschaftswissenschaft
 - Digital Humanities
- **Integration in die Informatik:**
Neben- bzw. Anwendungsfach für
 - B.Sc.: Informatik, Angewandte Informatik
 - M.Sc.: Informatik, Computational Science

Computerlinguistik in Jena (2/2)

- Aktive Forschergruppe
 - Lehrstuhl für Computerlinguistik = **Jena University Language & Information Engineering (JULIE) Lab**
 - Hohe internationale Visibilität (Publikationsdichte)
 - Deutsche Forschungsgemeinschaft (DFG)
 - Aktuell: (1/5) SFB 1076 **AquaDiva – Biodiversität in der Critical Zone**
 - Aktuell: (1/5) Graduiertenkolleg **Modell ‚Romantik‘ [Digital Humanities]**
 - Bundesministerium für Bildung & Forschung (BMBF)
 - Aktuell: (1/7/26) Nationale Förderinitiative „**Systemmedizin**“ (J–L–AC)
 - Frühere Projekte: Forschungs-Cluster **JenAge** – Nationaler Forschungskern, **StemNet**
 - Förderinitiativen der Europäischen Union
 - Frühere Projekte: **MANTRA (SA)**, **CALBC (SA)**, **BOOTStrep (STREP)**, ..
- Ausgründung von Start-up-Firmen
 - *Averbis, TexKnowlogy*
- **Jobs, Jobs, Jobs ...** etwa als studentische Hilfskraft
- **Themen, Themen, Themen ...** BA- oder MA-Arbeit, Dissertation

Weitere Veranstaltungen

- Seminar zu M-GSW-09
 - Machine Reading – WWW-skalierbares automatisches Textverstehen
 - Do, 16-18, Fürstengraben 1, SR 164

Kleiner Exkurs zum Thema “Wissenschaftliche Exzellenz”

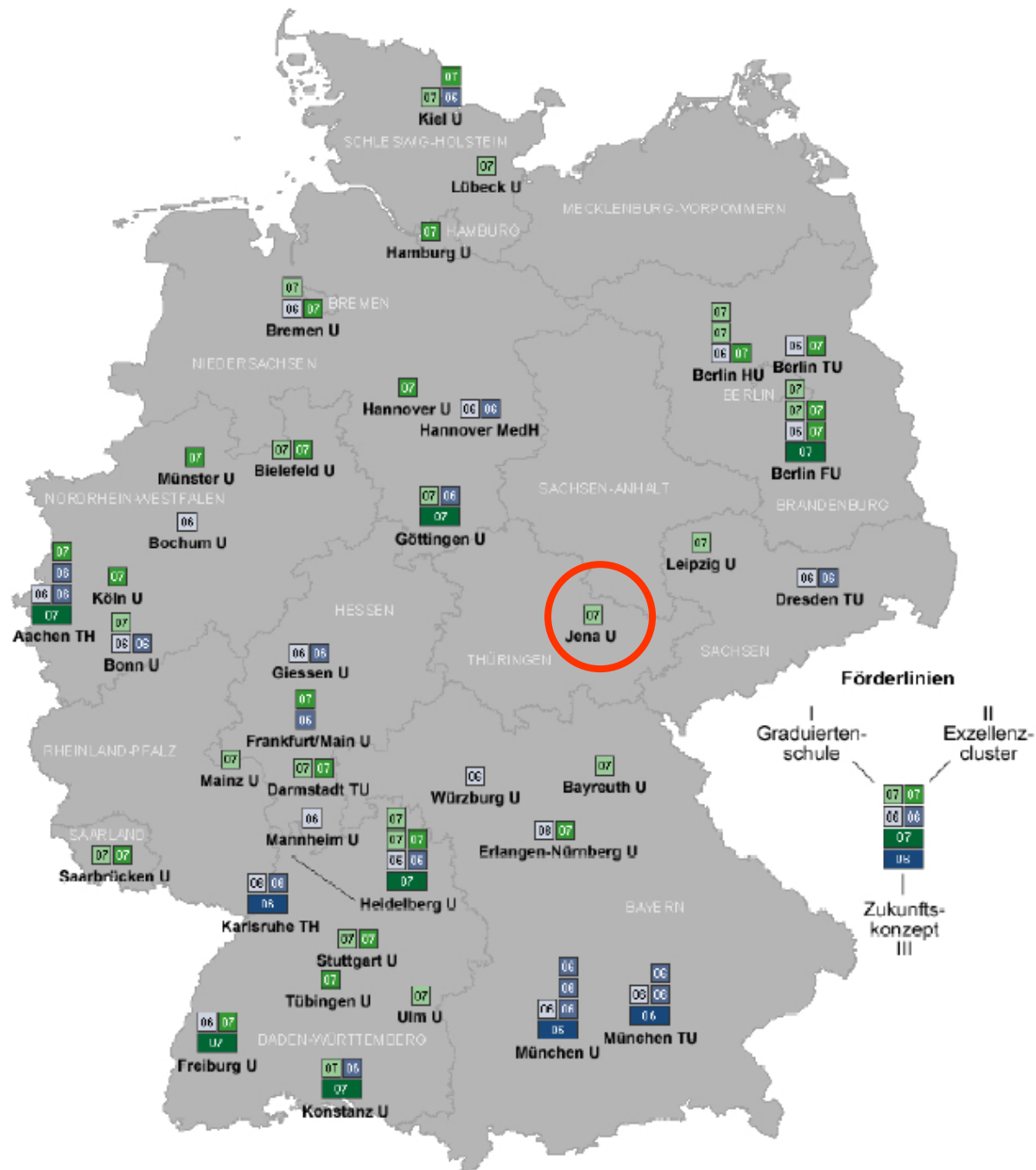
1. Exzellenzinitiative (2006-07)

87 deutsche Universitäten

44 in Förderlinien

FSU Jena: 1 Graduiertenschule:
Jena School for
Microbial Communication

9 Elite-Universitäten (I)
(FUB, FR, GÖ, HD, KA, KN,
MUM, TUM, RWTH AC)



2. Exzellenzinitiative (2010-12)

87 deutsche Universitäten

11+45+43 = 89 in Förderlinien

FSU Jena: 1 Graduiertenschule:
Jena School for
Microbial Communication

9 Elite-Universitäten (I)
(FUB, FR, GÖ, HD, KA, KN,
MUM, RWTH AA, TUM)

11 Elite-Universitäten (II)
(TUDD, FUB, HB, HUB, HD, K, KN,
MUM, TUM, RWTH AC, TÜ)



3. Exzellenzinitiative (2017-22)

Förderung der neuen Exzellenzcluster (EXC) ab 1. Januar 2019

Entscheidung der Exzellenzkommission vom 27. September 2018



Woher kommt Exzellenz ?

- (High-impact-)Publikationen
 - Wissenschaftspreise
 - Drittmiteleinwerbungen
 - SFBs, Graduiertenschulen ...
 - Zukunftsentwürfe
-
- Im internationalen Kontext weltweit sichtbar sein (visibility)

Ein Beispiel für den Nachweis wissenschaftlicher Exzellenz

- **Semantik**
 - Bedeutung von Sprache
- **Semantische Textanalytik**
 - Inhaltliche Analyse von Texten
- **Informationsbeschaffung für Biologen und Mediziner**
 - Medline/PubMed: mehr als 27M Dokumente
- **„Weltmeisterschaft“ für semantische Textanalytik**
 - Wo ist Jena (JULIE Lab) ?

<Semantische Textanalytik>

- **Natürlichsprachliche Semantik**
 - Lexikalische Semantik, Satzsemantik
- **Term-Semantik**
 - Termvarianten: Synonyme, Akronyme, Abkürzungen
- **Typen-Semantik**
 - Generalisierung auf Klassen
- **Propositionale Semantik**
 - Prädikationen : $p(a_1, \dots, a_n)$, a_i kann Term sein, aber auch eine Prädikation

Propositionale Semantik

Annotation Results for 7591091.xmi in C:\Users\jwermt\ Desktop\HahnBOOTStrip

Attenuation of **gamma interferon**-induced tyrosine phosphorylation in mononuclear phagocytes infected with *Leishmania donovani*: selective inhibition of signaling through Janus kinases and **Stat1**. The induction of gene transcription in response to **gamma interferon** is impaired in mononuclear phagocytes infected with *Leishmania donovani*, and the mechanisms involved are not fully understood. The changes in gene expression brought about by **gamma interferon** are thought to **involve** transient **increases** in the activities of cellular protein tyrosine kinases, including the Janus kinases **Jak1** and **Jak2**, **leading** to tyrosine phosphorylation of the transcription factor **Stat1**. To investigate the mechanisms accounting for the impaired responses to **gamma interferon**, a model system for examining overall changes in protein tyrosine phosphorylation, **activation** of **Jak1** and **Jak2** and **phosphorylation** of **Stat1** was developed in phorbol 12-myristate 13-acetate-differentiated U-937 cells. Analysis of whole-cell lysates by antiphosphotyrosine immunoblotting showed that incubation with **gamma interferon** brought about specific increases in phosphotyrosine labeling of several proteins. Increased labeling of these proteins occurred to similar extents in control cells and in cells that had been infected with *L. donovani* for 16 h. **Jak1**, **Jak2**, and **Stat1** were **immunoprecipitated** from control and interferon-treated cells, and tyrosine phosphorylation of these proteins, detected by antiphosphotyrosine immunoblotting was used to measure their activation. Tyrosine **phosphorylation** of **Jak1**, **Jak2**, and **Stat1** **increased** markedly, in a dose-dependent manner, in U-937 cells incubated with **gamma interferon**. In contrast, in cells infected with *L. donovani*, tyrosine **phosphorylation** of **Jak1**, **Jak2**, and **Stat1** was markedly **impaired**. This effect was dependent upon the duration of exposure to *L. donovani* and was maximal and complete at 16 h. Results similar to those observed with U-937 cells were also obtained with human peripheral blood monocytes. These findings indicate that infection of human mononuclear phagocytes with *L. donovani* leads to **impaired gamma interferon-mediated** tyrosine **phosphorylation** and selective effects on the **Jak-Stat1** pathway. Unresponsiveness to **gamma interferon** for activation of this pathway may explain impaired transcriptional responses in *leishmania*-infected cells.

Legend

☒ EventMention

☒ Gene

Propositionale Semantik

Annotation Results for 7591091.xmi in C:\Users\jwermt\Desktop\HahnBOOTStrip

Attenuation of **gamma interferon**-induced tyrosine phosphorylation in mononuclear phagocytes infected with *Leishmania donovani*: selective inhibition of signaling through Janus kinases and **Stat1**. The induction of gene transcription in response to **gamma interferon** is impaired in mononuclear phagocytes infected with *Leishmania donovani*, and the mechanisms involved are not fully understood. The changes in gene expression brought about by **gamma interferon** are thought to **involve** transient **increases** in the activities of cellular protein tyrosine kinases, including the Janus kinases **Jak1** and **Jak2**, **leading** to tyrosine phosphorylation of the transcription factor **Stat1**. To investigate the mechanisms accounting for the impaired responses to **gamma interferon**, a model system for examining overall changes in protein tyrosine phosphorylation, **activation** of **Jak1** and **Jak2** and **phosphorylation** of **Stat1** was developed in phorbol 12-myristate 13-acetate-differentiated U-937 cells. Analysis of whole-cell lysates by antiphosphotyrosine immunoblotting showed that incubation with **gamma interferon** brought about specific increases in phosphotyrosine labeling of several proteins. Increased labeling of these proteins occurred to similar extents in control cells and in cells that had been infected with *L. donovani* for 16 h. **Jak1**, **Jak2**, and **Stat1** were **immunoprecipitated** from control and interferon-treated cells, and tyrosine phosphorylation of these proteins, detected by antiphosphotyrosine immunoblotting was used to measure their activation. Tyrosine **phosphorylation** of **Jak1**, **Jak2**, and **Stat1** **increased** markedly, in a dose-dependent manner, in U-937 cells incubated with **gamma interferon**. In contrast, in cells infected with *L. donovani*, tyrosine **phosphorylation** of **Jak1**, **Jak2**, and **Stat1** was markedly **impaired**. This effect was dependent upon the duration of exposure to *L. donovani* and was **maximal and complete at 16 h**. **Results similar to those observed** with U-937 cells were also obtained with human peripheral blood monocytes. These findings indicate that infection of human mononuclear phagocytes with *L. donovani* leads to **impaired gamma interferon-mediated** tyrosine **phosphorylation** and selective effects on the **Jak-Stat1** pathway. Unresponsiveness to **gamma interferon** for activation of this pathway may explain impaired transcriptional responses in leishmania-infected cells.

Legend

☒ EventMention ☒ Gene

Click In Text to See Annotation Detail

Annotations

- EventMention
 - EventMention ("impaired")
 - begin = 1794
 - end = 1802
 - confidence = null
 - componentId = de.julielab.jules.JREX
 - id = E921
 - specificType = Negative_regulation
 - ref = null
 - resourceEntryList = null
 - textualRepresentation = null
 - arguments = FSArray
 - arguments = ArgumentMention ("phosphorylation")
 - begin = 1740
 - end = 1755
 - confidence = null
 - componentId = null
 - id = null
 - ref = EventMention ("phosphorylation")
 - begin = 1740
 - end = 1755
 - confidence = null
 - componentId = de.julielab.jules.JREX
 - id = E619
 - specificType = Phosphorylation
 - ref = null
 - resourceEntryList = null
 - textualRepresentation = null
 - arguments = FSArray
 - arguments = ArgumentMention ("Jak2")
 - begin = 1765
 - end = 1769
 - confidence = null
 - componentId = null
 - id = null
 - ref = Gene ("Jak2")
 - role = Theme

Challenge Competitions



- ParsEval, SemEval, RTE, ...
- MUC, ACE, TAC, SUMMAC
- BioCreative I, II. II.5, III, IV, LLL, NLPBA
- TREC (Genomics), CLEF eHealth, i2b2
- **BioNLP'09 Shared Task on Event Extraction**
 - <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/>
- CALBC, MANTRA

Challenge Competition (1/3)

1. (vertrauenswürdiger, fairer, objektiver)
Ausrichter konstituiert sich
 - Thematik des Challenge festlegen
 - Textauswahl, Formate etc.
 - Wettbewerbssoftware bereitstellen
2. Anfertigung des Goldstandards (*ground truth*)
 - Aufspaltung in
 - Training-Set (70/90)
 - Test-Set (30/10)

Challenge Competition (2/3)

3. Freigabe des Training-Set (Dauer: 3-6 W)

- Teilnehmer trainieren ihr System am Training-Set
- Vergleich eigener Ergebnisse gegen Goldstandard
- Teilnehmer fixiert am Ende der Trainingsphase n optimale Systemzustände (*frozen system*)

4. Freigabe des Test-Set (Dauer: 2-3 T)

- Frozen system operiert auf Test-Set

Challenge Competition (3/3)

5. Abgabe der Ergebnisse beim Ausrichter
6. Auswertung der Ergebnisse des Test-Set-Laufs beim Ausrichter
 - Vergleich eigener Ergebnisse gegen Goldstandard
 - Standardisierte Metriken für Qualitätsmessung (precision, recall, F-score)
7. Vergleich und Ranking aller Teilnehmer durch Ausrichter
 - anonym (bei Bedarf)

And the winner is ...

Final Evaluation Results on ALL-TOTAL events by Approximate Span & Recursive Matching

<u>Team</u>	<u>gold (match)</u>	<u>answer (match)</u>	<u>recall</u>	<u>precision</u>	<u>fscore</u>
U Turku (FIN)	3182 (1487)	2541 (1486)	46.73	58.48	51.95
FSU Jena/JULIELab (GER)	3182 (1458)	3068 (1458)	45.82	47.52	46.66
Concordia U/CLaC (CAN)	3182 (1113)	1807 (1113)	34.98	61.59	44.62
U Tokyo+DBCLS (JAP)	3182 (1174)	2110 (1173)	36.90	55.59	44.35
Ghent U/VIB(BEL)	3182 (1063)	2062 (1063)	33.41	51.55	40.54
U Tokyo/Tsujii Lab (JAP)	3182 (895)	1671 (895)	28.13	53.56	36.88
U New South Wales (AUS)	3182 (898)	1957 (896)	28.22	45.78	34.92
U Zurich (SWI)	3182 (883)	1895 (883)	27.75	46.60	34.78
Arizona SU+HUB+BU(USA)	3182 (688)	1106 (688)	21.62	62.21	32.09
U Cambridge (UK)	3182 (672)	1181 (672)	21.12	56.90	30.80
U Antwerp/CNTSLTG (BEL)	3182 (716)	1501 (716)	22.50	47.70	30.58
U Manchester (UK)	3182 (702)	1444 (702)	22.06	48.61	30.35
SCAI Fraunhofer Inst (GER)	3182 (826)	2278 (826)	25.96	36.26	30.26
UAveiro (POR)	3182 (666)	1351 (666)	20.93	49.30	29.38
Team 24 (???)	3182 (722)	1778 (721)	22.69	40.55	29.10
U Szeged (HUN)	3182 (685)	1852 (685)	21.53	36.99	27.21
NICTA/U Melbourne (AUS)	3182 (555)	1388 (555)	17.44	39.99	24.29
CNB Madrid (ESP)	3182 (911)	4362 (911)	28.63	20.88	24.15
U Colorado/BTMG (USA)	3182 (428)	596 (428)	13.45	71.81	22.66
Arizona SU/CIPS (USA)	3182 (725)	3809 (725)	22.78	19.03	20.74
U Michigan (USA)	3182 (968)	6859 (968)	30.42	14.11	19.28
Sirma/Ontotext (BUL)	3182 (358)	538 (358)	11.25	66.54	19.25
Team 09 (???)	3182 (372)	1184 (372)	11.69	31.42	17.04
KoreaU (KOR)	3182 (299)	485 (299)	9.40	61.65	16.31

Post-competition Results I

Final Evaluation Results on ALL-TOTAL events by Approximate Span & Recursive Matching

<u>Team</u>	<u>gold (match)</u>	<u>answer (match)</u>	<u>recall</u>	<u>precision</u>	<u>fscore</u>
U Turku (FIN)	3182 (1487)	2541 (1486)	46.73	58.48	51.95
FSU Jena/JULIELab (GER)	3182 (1458)	3068 (1458)	45.82	47.52	46.66

Evaluation Results on ALL-TOTAL events by Approximate Span & Recursive Matching
after System Overhaul and further Tuning

<u>Team</u>	<u>gold (match)</u>	<u>answer (match)</u>	<u>recall</u>	<u>precision</u>	<u>fscore</u>
U Turku (FIN)					52.86
FSU Jena/JULIELab (GER)					51.10

In:
Computational Intelligence
Vol. 27, 2011, No.4, pp.610-44.

Post-Competition Results II

Participant	Rank in F1 score	Total			Localization	Binding	Gene expression	Transcription	Protein catabolism	Phosphorylation	Regulation	Positive Regulation	Negative Regulation	BioNLP '09 ST Total Evaluation
		F1	PR	RC	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1
JULIE Lab JReX [1]	1	51.09	57.69	45.85	61.60	49.24	72.48	42.99	80.00	81.99	31.20	40.39	38.47	46.66
UTurku [2]	2	49.91	56.32	44.81	55.85	45.43	71.67	50.21	50.00	79.70	33.97	38.66	36.28	51.95
EventMine [3]	3	48.20	64.00	38.65	63.20	39.86	72.63	50.00	60.87	81.29	28.77	28.25	32.62	36.88
BExtract [4]	4	44.48	61.56	34.82	51.45	26.97	65.14	24.71	60.00	80.69	32.21	35.83	33.27	44.62
VIBGhent [7]	5	42.44	59.05	33.12	51.79	34.42	69.57	57.14	68.97	76.23	19.39	23.34	26.67	40.54
TheBeast [8]	6	37.19	48.15	30.30	48.98	34.50	59.28	17.48	72.00	72.79	29.96	29.57	27.32	44.35
UMich [9]	7	36.34	35.57	37.15	53.47	31.75	66.00	30.06	58.06	77.15	14.29	21.50	26.61	19.28
Moara [6, 10]	8	29.50	31.99	27.31	44.19	28.36	58.79	26.40	50.00	52.88	10.83	14.68	13.16	24.15
CCP-BTMG [13]	9	22.03	70.03	13.07	17.80	20.92	51.07	22.93	40.00	33.33	5.79	6.69	4.01	22.66

In:
BMC Bioinformatics
 Vol. 12, 2011, No.481

Merkmale von Challenge Competitions

- Internationaler Ideen-Wettbewerb
- Intersubjektive Bewertung
- Saubere Vergleichsmaßstäbe: Metriken
- Experimente
- Trennung Experimentator/Entwickler
- „sportlicher“ Aspekt
- Offenlegung der Methoden
 - Treiber für Methodenfortschritt
 - Latente Gefahr des Methodenkonservatismus‘
- Exemplarischer Fall empirischer Wissenschaft: Rationalität

</Exkurs>

Computerlinguistik I

- Linguistik: Gegenstandsbereich sind (überwiegend) **natürliche Sprachen**
 - Deutsch, Englisch, Französisch, ...
- Beispiele für **formale Sprachen**
 - $L = \{a^n b^n, n \in \mathbb{N}\}$
= {ab, aabb, aaabbb, aaaabbbb, ... }
 - jede Programmiersprache, Auszeichnungssprache
 - JAVA, C++, ..., XML, HTML, ...
 - jede Logik
 - Aussagenlogik, Prädikatenlogik, Typenlogik, ...
 - Differentialgleichungen, Integrale, Vektoren, ...

Formale Sprachen

- **Konstruiert**
 - Rein definitorischer (konstruktiver) Ansatz
- **Möglichst non-ambig**
 - Eindeutige syntaktische wie semantische Strukturen
- **Statisch**
 - zum Definitionszeitpunkt komplett fixiert
 - Endliches Vokabular
- **„Einfache“ Beschreibung**
 - Wenige Regeln, wenige Axiome
 - meist wenige Elemente umfassendes Vokabular („Lexikon“)
 - Wenige Schichten: Syntax, Semantik; keine Pragmatik
- **striktes Wohlgeformtheitskriterium**
 - Außer-definitorische Strukturen sind nicht wohlgeformt
 - ... und damit nicht prozessierbar

Natürliche Sprachen

- Konventionalisiert durch ‚sozialen Vertrag‘ einer Sprechergemeinschaft
 - Ausübung des Sprechens unterliegt sozialen Normen, Gewohnheiten und (impliziten) Übereinkünften (Regelkonformität)
- Hochgradig ambig
 - Mehrdeutige lexikalische, syntaktische, semantische, pragmatische Strukturen
- Dynamisch
 - Sprache verändert sich im Laufe der Zeit (Lexikon, Syntax)
 - Unendliches Vokabular (Komposition, Derivation)
- Komplexe Beschreibungen
 - Viele Regeln, viele Axiome
 - Sehr großes Vokabular („Lexikon“)
 - Starke Schichtung von Beschreibungsebenen
- laxe Wohlgeformtheitskriterien
 - Außer-definitorische Strukturen sind zwar nicht wohlgeformt, werden aber (bis zu einem gewissen Grad) verstanden