Motivation und Ressourcen
Mathe III

Prof Dr. Matthew Crocker

Universität des Saarlandes

28 Mai 2015

Wichtige Informationen

Finden Sie hier:
http://www.coli.uni-saarland.de/∼crocker/mathe3/mathe3.html

- Folien, Übungsmaterialien (Aufgabenblätter, Korpora etc.)
- Kontakt (crocker@coli.uni-sb.de, Büro: C7 1, Raum 1.15)

Außerdem:

- 4+2 SWS / 9 credits für 6h Präsenz & 12h "Eigenstudium"
- Voraussetzung für Schein: mind. 50% in Übungsblättern
- Für Tutorien ist ein Coli-account notwendig
  ($http://www.coli.uni-saarland.de/sg/application\_for\_userid.pdf$)

Übersicht

Date	Montag 4-6pm	Date	Dienstag 4-6pm	Date	Donnerstag 2-4pm
15/06/15	No Lecture	16/06/15	Zwischenklausur	18/06/15	L2: HMMs (Ü1 austeilen)
22/06/15	L3: Entropy (Ü2 austeilen)	23/06/15	Corpora (Tut1)	25/06/15	L4: Kollocations (Ü3 austeilen)
29/06/15	L9: PCFGs 1 (Vg + Ü4 austeilen)	30/06/15	Entropy (Tut2)	02/07/15	L5: Psycholinguistics
06/07/15	L6: Decision Trees	07/07/15	Kollocations (Tut3)	09/07/15	L7: Naive-Bayes (Ü5 austeilen)
13/07/15	L8: Clustering	14/07/15	PCFG-Ü besprechen (Tut4)	16/07/15	L10: PCFGs 2
20/07/15	Probeklausur, Abgabe PCFGs Ü	21/07/15	Fragen, PK Solution, Besprechung Zwischenklausur EL	23/07/15	Decision Trees (Tut5)
27/07/15	Free	28/07/15	Klausur	30/07/15	Free

Natürliche (menschliche) Sprache

Zitat

. . . language is **a biological system**, *and biological systems typically are* **"messy"**, *intricate,* **the result of evolution***ary "tinkering", and* **shaped by accidental circumstances** *and by . . . conditions that hold of complex systems. . .*

(Chomsky, *The minimalist program*)

Natürliche (menschliche) Sprache

Aus dem Verbmobil-Korpus

Spontan-sprachliche Terminabsprache Deutsch-Englisch-Japanisch:

. . . bei mir ist die Woche davor schlecht, **also**, *die Woche nach Pfingsten*, **und** *die erste Maiwoche,* **also,** *alles andere*
wäre stünde *zur Disposition, dann würde ich mal sagen, dass wir den ersten Termin auf Montag, den neunten Mai legen. . .*

Kompetenz und Performanz

Kompetenz

- Potenzielle, idealistische (angeborene) Fähigkeit zur Sprache bzw. Wissen um die Sprache
- Endliche Menge von Sprachregeln, die Sprecher verinnerlicht haben und die zum Verstehen und Produzieren von Sprache dienen
- Beschreibt die wohlgeformten Äusserungen einer Sprache
- Kann man nicht direkt beobachten

Performanz

- Anwendung der zur Kompetenz gehörenden Regeln
- Tatsächlich vorkommende Äusserungen
- Zu beobachtendes Verhalten

Zwei Ansätze

The Armchair Linguist

He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting "Wow, what a neat fact", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

(Charles Fillmore)

Zwei Ansätze

The Corpus Linguist

He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first words of a sentence versus as the second word of a sentence.

(Charles Fillmore)

Regelbasierte Modelle

- Modellierung durch theoretische Überlegung
- Gesucht werden Regeln,
  - die alle Fälle eines Phänomens erfassen,
    aber nicht übergenerieren
  - die einfach genug sind, um von einem Computer berechnet zu werden
    (kein Rückgriff auf Weltwissen usw.)

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Regelbasierte Modelle

Bsp.: Woran erkenne ich ein Adjektiv?

Ich möchte Ihnen für den Bericht über den **siebenten** *Bericht über* **staatliche** *Beihilfen in der* **europäischen** *Union danken.*

(European Parliament Proceedings) Es ist schwer, korrekte und vollständige Regeln zu schreiben

- Regel 2 ist zu liberal (möchte = Adj)
- Regel 3 ist zu streng (staatliche = NAdj)
- Das System trifft eine harte Entscheidung für jede Instanz
- Keine Möglichkeit, über "Wahrscheinlichkeit" zu sprechen

Regelbasierte Modelle

- Erfolgreich für Morphologie, Grammatiken (Grammatiktheorie), formale semantische Analyse
- Vorteile
  - Erlaubt Modellierung komplexer Phänomene ("tiefe" Analyse)
  - Kann negative Evidenz einbeziehen (=Was **nicht** möglich ist)
  - Ergebnis ist für Menschen verständlich
  - Bietet oft eine Erklärung des Phänomens an

Regelbasierte Modelle

Nachteile regelbasierter Systeme:

- Nicht geeignet für stetige Phänomene
- Können keine Präferenzen ausdrücken
- Häufig präskriptiv statt deskriptiv
- Mangel an Robustheit: Schon bei kleinen Fehlern in der Eingabe bricht die Analyse ab
- Objektivität?
- Hand-Arbeit: Hoher Aufwand
  Die English Resource Grammar (ERG) wird seit Mitte der 90er Jahre in mehreren grossen CL-Projekten entwickelt, aber es wird noch daran gearbeitet!

Korpuslinguistik und Statistik

- Daten-orientierte Untersuchungen:
  Modellierung durch Sichtung von Beispielen
- Erkennung ähnlicher Muster und
  Regelmässigkeiten in den Daten
- Vorteile
  - Auf Grund von Daten trainiert: Weniger Handarbeit
    (Einsatz maschineller Lernverfahren)
  - Bestimmung der wahrscheinlichsten Lesart
  - Robust: Können mit fehlerhafter oder unbekannter
    Eingabe umgehen
  - Modelle können Übergenerierung erlauben, um Robustheit zu erreichen
  - Zugriff auf in den Daten implizites Weltwissen
  - Schnelle Modellierung neuer Domänen, Sprachen, usw.

Korpuslinguistik und Statistik

Einige Beispiele

- Lexikalische Präferenzen
  - Wortkategorie: *bank* = Substantiv 85 %, Verb 15 %
  - Bedeutung: *bank* (river) = 22 %, *bank* (money) = 78 %
- Syntax:
  - realized + NP = 20 %
  - realized + S = 65 %
  - realized + other = 15 %
- Anaphern: *He* bezieht sich auf Englisch in 63 % der Fälle auf das Subjekt des vorigen Satzes
- Textanalyse: Autor X verwendet das Wort *bezüglich* "signikant" öfter als Autor Y

Korpuslinguistik und Statistik

- Nachteile
  - Flache Analyse (Engl. ,,shallow")
  - Modelle nur approximativ richtig
  - Schwierige Probleme können oft nicht zuverlässig modelliert werden
  - Modelle für Menschen schwierig zu verstehen und abzuändern
  - Rein descriptiv, keine Erklärung
  - Abhängigkeit von den Daten
  - Problem mit unbekannten Wörtern/Strukturen (Sparse Data)
- Erfolgreich für:
  - Wortartenanalyse
  - Automatische syntaktische Analyse

Parallel mit Nativismus vs. Empirie

Welche Rolle spielt Spracherfahrung beim Sprachenlernen?

- Nativismus: Sprache ist sehr komplex, daher muss die Fähigkeit dazu und deren Grundprinzipien beim Menschen angeboren sein

  (Vgl. Chomsky's *Principles and Parameters*:
  - Sowohl Prinzipien als auch Parameter sind Sprachuniversalien
  - Menschen kennen die Prinzipien von Geburt an, z. B. dass alle Sätze ein Subjekt haben, auch wenn es in manchen Sprachen overt (=sichtbar) weggelassen werden kann
  - Spracherwerb besteht darin, die Parameter für die eigene Muttersprache zu setzen: SVO oder OVS? usw.)
- Empirizismus: Sprachliches Wissen erwerben Kinder ausschliesslich durch das Hören der Sprache ihrer Eltern

Sprachressourcen

Überblick

- Korpora (Singular Korpus, neutrum!):
  Textkollektionen (z.T. mit zusätzlichen Informationen angereichert)
- Wörterbücher, Lexika, Thesauri, manche Enzyklopädien
- Ontologien, semantische Netze und sonstige Formen von Wissensrepräsentation

Korpora

Definition

- *Ein Korpus (n.!) ist eine endliche Sammlung von konkreten sprachlichen Äusserungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen*

(Lexikon der Sprachwissenschaft)

- *Eine idealerweise repräsentative, möglicherweise auf einem Bereich eingeschränkte Sammlung von Texten einer gegebenen Sprache, die zum Zwecke linguistischer Analyse zusammengestellt wurde*

(Francis, 1964)

(Francis and Kucera: Ersteller des Brown Corpus, frühes Korpora fürs Englische)

Aufbereitung: Rohe Korpora

- Die grössten Korpora sind rohe Korpora (heute: das Internet selbst)
- Einsatz in der Lexikographie:
  Manuelle Sichtung Beispiele (Konkordanz), um Wortbedeutungen zu bestimmen, sowie Neologismen und Kollokationen zu entdecken

hängen , Packpferde mit Brennholz ; Frauen	backen	Brot , Kinder hüten Ziegen . Von Zeit zu Zeit
unmusikalisch . Aber sie kann Pfannkuchen Brot ,	backen	Nun folgt die konkrete Utopie (oder was m
Bei 170 Grad , Gas : Stufe 3 etwa 1 1/4 Std.	backen	. Vor dem Herausnehmen erkalten lassen . R
Leute an . Lasst uns anfangen , ich muss Brot	backen	, meinte er unwirsch und genehmigt sich un
kann doch nicht jeder seine eigenen Brötchen .	backen	, mahnte Scherf . Dann wieder Fragen : Ob
e , und die zieht er formvollendet durch : Wir	backen	einen guten Kurzlm . An der Idee blieb auc
ssen . In heissem Backfett kleine Pfannkuchen	backen	und mit saurer Sahne und Kaviar servieren .
zu besticken , Kaffee zu kochen und Kekse zu	backen	, um so ihrer Verpichtung gegenüber dem
. Im Moment aber muss er ganz kleine Brötchen	backen	. Der Grüne sieht sich einer erdrückenden sc
, 1/2 Stunde ziehen lassen , dann goldbraun	backen	. Mit Erdbeeren garnieren . Alle Rezepte aus
Halloween höhlen sie einen Kürbis aus und	backen	Pumpkin-Pie . Die Prices sind eine durchsch
ade oder Quark . Schwaben südlich der Donau	backen	Brot , wie die riesigen Knauzawecka , noch
schwimmen gehen , nachtwandern , Stockbrot	backen	, die Bauern besuchen , basteln , spielen . Bei

Aufbereitung: Formatierung

- Alte Korpora: Ad-hoc Format
- Interlinear format (hier: Wort_PoS_Lemma):
  `John_PN_john left_VBP_leave ._PUNC_period`
- Spalten (Susanne, 1. Spalte: Satz- und Wort-Id)

A12:0210	John	john	PN
A12:0211	left	leave	VBN
A12:0212	.	Period	PUNC

- SGML Mark-up (veraltet, Vorgänger von XML)
- Penn Treebank (syntaktische Annotation)

```
( (S
    (PP-LOC (IN In)
      (NP
        (NP (DT an) (NNP Oct.) (CD 19) (NN review) )
        (PP (IN of)
          (NP (`` ``)
            (NP-TTL (DT The) (NN Misanthrope) )
```

- Heute: meist XML (Vorteil: Allgemeine Tools)

Beispiel aus dem BNC (SGML)

```
<s n=0001>
 <w NN1>INTRODUCTION
</head>
<p>
<s n=0002>
 <w AT0>The <w AJ0>extensive <w NN1>upland <w NN2-VVZ>landscapes <w PRF>of
 <w AT0>the <w NP0>UK<c PUN>, <w CJC>and <w AT0>the <w AJ0>varied <w CJC>and
 <w AJ0>rich <w NN1>wildlife <w PNP>they <w VVB>support<c PUN>, <w VBB>are <w AT0>the
 <w NN1>product <w PRF>of <w NN2>centuries <w PRF>of <w AV0>predominantly
 <w AJ0>pastoral <w AJ0-NN1>agricultural <w NN1>activity<c PUN>.
<s n=0003>
 <w PRP>In <w AT0>the <w AJ0-NN1>past<c PUN>, <w AT0>the <w NN1>use <w PRF>of
 <w DT0>these <w NN2>uplands <w PRP>for <w NN0>sheep <w CJC>and <w NN1>beef
 <w NN2>cattle <w NN1-VVG>rearing <w VHZ>has <w XX0>not <w VVN>conflicted
 <w AV0>significantly <w PRP>with <w AT0>the <w NN1>need <w TO0>to <w VVI>retain
 <w NN2>habitats <w PRP>such as <w NN2>moorlands<c PUN>, <w NN1>hill
 <w NN2>grasslands<c PUN>, <w AJ0>high <w NN1>altitude <w AJ0>montane
 <w NN1>vegetation<c PUN>, <w AJ0-VVD>enclosed <w NN2>pastures <w CJC>and
 <w NN1-VVB>hay <w NN2>meadows<c PUN>, <w NN2>wetlands <w CJC>and <w AJ0>native
 <w NN2>woodlands<c PUN>, <w DTQ>which <w VVB>form <w AT0>the <w NN1>basis <w PRF>of
 <w AT0>the <w NN1>nature <w NN1>conservation <w NN1>interest <w PRF>of <w AT0>the
 <w CRD>9.68 <w CRD>million <w NN2>hectares <w PRF>of <w NN1>upland <w PRP>in
 <w AT0>the <w NP0>UK<c PUN>.
```

Entwicklung von Sprachressourcen

- Roher Text genügt oft nicht, daher wird es ergänzt um Satzgrenzen, Wortkategorien,. . .
- Korpus-Annotation macht enthaltene linguistische Information explizit, kann aber falsch sein

Prinzipien für Korpus-Annotation (Leech, '93)

- Sowohl Annotation als originales (rohes) Korpus sollte von einander trennbar sein
- Die Annotation sollte theorieunabhängig und neutral sein
- Die Annotationsmethode (manuell, maschinell, oder Kombination davon) sollte bekannt sein
- Die Annotationsrichtlinien sollten mit allen Details verfügbar sein

Entwicklung von Sprachressourcen

- Roher Text genügt oft nicht, daher wird es ergänzt um Satzgrenzen, Wortkategorien,. . .
- Korpus-Annotation macht enthaltene linguistische Information explizit, kann aber falsch sein

Prinzipien für Korpus-Annotation (Leech, '93)

- Sowohl Annotation als originales (rohes) Korpus sollte von einander trennbar sein
- Die Annotation sollte theorieunabhängig und neutral sein
- Die Annotationsmethode (manuell, maschinell, oder Kombination davon) sollte bekannt sein
- Die Annotationsrichtlinien sollten mit allen Details verfügbar sein

Aufbereitung: Annotation

Annotation: Hinzufügen von Information

Probleme:

- Welcher TAGset, welcher (Grammatik-) Formalismus, . . . ?
- Interpretation (HPSG/LFG vs. funktionale Grammatik)
- Wegen Ambiguität ist Annotation nicht einfach
- **Manuelle** Annotation
  - Annotationsaufwand für ein Wort: 30 Sekunden
  - 1M Worte: 500 000 Minuten = 5 Jahre
  - Plus Aufwand fuer Qualitätssicherung
  - Fehler und Inkonsistenz nicht vollständig vermeidbar
- **Automatische** Annotation macht systematische Fehler

Annotation: Korrektheit

- Wichtigstes Kriterium: Korrektheit
- Falsche Annotation führt zu falschen Modellen
- Manuelle Annotation
- Selbst manuelle Annotation ist nie fehlerfrei
  - Grund 1: Unaufmerksamkeit der Annotatoren
  - Grund 2: Schwierigkeit der Aufgabe
  - Es ist schwierig, über grosse Textmengen konsistent zu sein (intra-Annotatoren agreement)
  - Verschiedene Leute können systematisch ein Phänomen anders empfinden und annotieren (inter-Annotatoren agreement)

Mögliche Lösungen

- Annotationsmöglichkeiten gering halten (z. B. kleines Tagset), um schwierige Entscheidungen aus dem Weg zu gehen
- Bei Unsicherheit mehrere Tags zuweisen (dokumentieren, dass es Unsicherheit gab)

  z. B.: "Ambiguity Tags" im BNC:
  AJ0-AV0 (Adjectiv oder Adverb), mit Präferenz für AJ0
- Automatische Annotation mit der Überprüfung durch menschliche Annotatoren kombinieren
- Bootstrapping:
  Annotated corpora used to train & improve the annotation tools

Merkmale von Korpora

- Sprache: monolingual vs. bilingual vs. multilingual; vergleichbar vs. parallel, aligniert
- Textart, Inhalt, Genre, Domäne:
  - Spontansprache: Usenet, Wizard-of-Oz Experimente
  - Editiert: Zeitungsartikel, Romane, Fachtexte, Lyrik,...
  - Ausgewogenheit: homogen vs. heterogen, unbalanciert vs. balanciert
- Geschriebene Sprache vs. gesprochene Sprache
- Umfang (Tokens, Types), Zeitraum
- Format, Text oder Binär (indexiert)
- Medium (Text, Audio, Transkripte, Video, usw.)
- Aufbereitung und Annotation
- Urheber- und Nutzungsrechte, Preis
- Standard-Referenz: Allgemeine Verfügbarkeit

Überblick über Sprachressourcen

Für die unterschiedlichen Aufgaben, mit denen sich die CL beschäftigt wurden unterschiedliche Korpora gesammelt / annotiert:

Typ der Annotation	Corpus
roh	Gigaword (1.8 Millarden Worte) TAZ corpus (Deutsch)
Part-of-Speech TAGs	British National Corpus (BNC), 100M Worte American National Corpus (ANC), 22M Worte Huge German Corpus (HGC), 200M Worte
Satzstruktur (Baumbanken)	Penn Treebank (1M Worte, Englisch) NEGRA und TIGER (70.000 Sätze, Deutsch) Prague Dependency Treebank (Czech) weitere Sprachen: Französisch, Chinesisch...
Semantische Rollen	PropBank (Englisch) SALSA (Deutsch)
Diskursrelationen	Penn Discourse Treebank

(Achtung: keine vollständige Liste! nur Beispiele!)

Überblick über Sprachressourcen (2)

Korpora mit unterschiedlichen Modalitäten:

Annotationsart	Corpus
Spoken Language	Christine (200,000wds)
Dialog	MapTask (Scottish English) Communicator Corpus
Meetings (multimodal)	AMI Transkipt von Besprechungen
Übersetzung	Crater corpora (vglb. Daten) Parallel: Hansard Corpus, EUROPARL
Eye-movement	Dundee Corpus
5grams	Google 5gram Corpus (1 trillion word tokens from Web)

Entwicklung von Korpora

Annotation: PoS-Tagging

- Manuell oder automatisch?
- Tag Sets sind unterschiedlich gross; sie varieren in sowohl innerhalb als auch unter Kategorien in ihrer Granularität

	Brown	Penn	Claws 1–8	STTS
Grösse	77/177	45	60–160	54

- Sie sind sprachspezifisch
- Manchmal enthalten sie Seltsamkeiten
  - Brown:
    VBG für Present Participles und für Gerunde
    John is purchasing apples
    The Fulton County purchasing department
  - Penn:
    TO sowohl für Präpositionen als auch vor Infinitiven
    (*I want to go to the store*)

Brown Tagset

–	dash
,	comma
:	colon
.	sentence closer (. ; ? *)
(	left paren
)	right paren
*	not, n't
ABL	pre-qualifier (quite, rather)
ABN	pre-quantifier (half, all)
ABX	pre-quantifier (both)
AP	post-determiner
AT	article (a, the, no)
BE	be
BED	were
BEDZ	was
BEG	being
BEM	am
BEN	been
BER	are, art
BEZ	is
CC	coordinating conjunction
CD	cardinal numeral
CS	subordinating conjunction
DO	do
DOD	did
DOZ	does
DT	sg. determiner (this, that)
DTI	sg. or pl. det. (some, any)
DTS	pl. determiner (these, those)
DTX	double conjunction (either)
EX	existential there
FW	foreign word
HV	have
HVD	had (past tense)
HVG	having
HVN	had (past participle)
HVZ	have, pres., 3rd p. sg.
IN	preposition
JJ	adjective
JJR	comparative adjective
JJS	semantic superl. adj. (chief, top)
JJT	superlative adjective
MD	modal auxiliary
NC	cited word
NN	singular or mass noun
NNS	plural noun
NP	proper noun
NPS	plural proper noun
NR	adverbial noun
OD	ordinal numeral
PN	nominal pronoun
PP$	determiner, possessive
PP$$	pronoun, possessive
PPL	sg. reflexive pers. pron.
PPLS	pl. reflexive pers. pron.
PPO	personal pronoun
PPS	3rd p. sg. nom. pron.
PPSS	other nominative pers. pron.
QL	qualifier (very, fairly)
QLP	post-qualifier (enough, indeed)
RB	adverb
RBR	comparative adverb
RBT	superlative adverb
RN	nominal adverb (here, indoors)
RP	particle (about, off, up)
TO	to (before infinitive)
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	pres. part./gerund
VBN	verb, past part.
VBZ	verb, 3rd p. sg. pres.
WDT	wh- determiner
WPO	wh- pronoun, object
WPS	wh- pronoun, nom.
WQL	wh- qualifier (how)
WRB	wh- adverb

BNC Tagset, CLAWS 5 or C5

AJ0	Adjective
AJC	Comparative adjective
AJS	Superlative adjective
AT0	Article
AV0	Adverb
AVP	Adverb particle (e.g. up, off, out)
AVQ	Wh-adverb
CJC	Coordinating conjunction
CJS	Subordinating conjunction
CJT	that
CRD	Cardinal number
ORD	Ordinal numeral
DPS	Possessive determiner or pronoun
DT0	General determiner-pronoun
DTQ	Wh-determiner-pronoun
EX0	Existential there
NN0	Common noun, neutral for number
NN1	Singular common noun
NN2	Plural common noun
NP0	Proper noun
PNI	Indefinite pronoun
PNP	Personal pronoun
PNQ	Wh-pronoun
PNX	Reflexive pronoun
POS	The genitive marker 'S or '
PRF	The preposition OF
PRP	Preposition, except OF
PUL	Punctuation: left bracket
PUN	Punctuation: general
PUQ	Punctuation: quotation mark
PUR	Punctuation: right bracket
TO0	Infinitive marker TO
UNC	Foreign words
VBB	The present tense of BE, except is
VBD	The past tense of BE
VBG	The -ing form of BE
VBI	The infinitive of BE
VBN	The past participle of BE
VBZ	IS, 'S
VDB	The finite base form of DO
VDD	The past tense of DO
VDG	The -ing form of DO
VDI	The infinitive of DO
VDN	The past participle of DO
VDZ	The -s form of DO
VHB	The finite base form of HAVE
VHD	The past tense of HAVE
VHG	The -ing form of HAVE
VHI	The infinitive of HAVE
VHN	The past participle of HAVE
VHZ	The -s form of HAVE
VM0	Modal auxiliary verb
VVB	The finite base of lexical verbs
VVD	The past tense of lexical verbs
VVG	The -ing form of lexical verbs
VVI	The infinitive of lexical verbs
VVN	The past participle of lexical verbs
VVZ	The -s form of lexical verbs
XX0	NOT or N'T
ITJ	Interjection
ZZ0	Alphabetical symbols

Überblick über Sprachressourcen

Syntax-Korpora ("Baumbanken")

- Penn Treebank: 1M Worte aus dem Wall Street Journal
- Deutsch:
  - NEGRA
    (20.000 Sätze Frankfurter Rundschau, 400K Worte)
  - TIGER
    (50.000 Sätze Frankfurter Rundschau = 1M Worte)
- Prague Dependency Treebank (Czech)
- Neuerdings auch für viele andere Sprachen:
  Chinesisch, Französische, usw.

Überblick über Sprachressourcen

NEGRA

- Als SQL Datenbank gespeichert;
  kann man in Bäume umwandeln
- Annotation:
  - PoS-tagged
  - Morphologische Annotation (60K)
  - Grammatische Funktionen
- Vorgehen:
  - Kombination aus automatischer Analyse und menschlicher Arbeit
  - Abfrage mit speziell dafür entwickelte Tools

Überblick über Sprachressourcen

Semantik-Korpora

[Peter]	Agent
gibt	
[Maria]	Recipient
[ein Buch]	Theme

- Satzteilen werden semantische Rollen zugeordnet
- Einsatz: Training semantischer Parsern
- Korpora:
  - Englisch: PropBank, auf Grundlage der Penn Treebank
  - Deutsch: SALSA, auf Grundlage von TIGER

Diskurs-Korpora

[Peter ist müde]. Grund
Deshalb [schläft er]. Folge

- Ordne Paaren von Sätzen Diskursrelationen zu:
  z. B. Begründung (weil), Zweck (damit),. . .
- Training von "Diskurs-Parsern"
- Korpora:
  DiscourseBank, auf Grundlage der Penn Treebank

Überblick über Sprachressourcen

Bilinguale Korpora

- Vergleichbare Daten:
  Crater corpora (English, French, Spanish)
- Parallel: Hansard Corpus, EUROPARL

Keine Korpora verfügbar

- Pragmatik:
  Intention der Sprecher, "was wirklich gemeint ist"
- Viele andere Sprachen, besonders für höhere Ebenen

Workbenches & Toolkits

Statistische Methoden: Sprachressourcen

- IMS Corpus Workbench: Aufbereitung und Abfrage
  `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
- Language Technology Group (LTG), Edinburgh
  `http://www.ltg.ed.ac.uk/`
- CMU Statistical Language Modeling Toolkit: Unix Tools
  `http://www.speech.cs.cmu.edu/speech/SLM_info.html`

Workbenches & Toolkits

Organisationen

- Linguistic Data Consortium (LDC)
  `http://www.ldc.upenn.edu/`
- Institut für deutsche Sprache (IDS), Mannheim
  `http://www.ids-mannheim.de/`
- European Language Resources Association (ELRA)
  `http://www.elra.info/`
- European Network of Excellence (ELSNET)
  `http://www.elsnet.org/`
- Lee Corpora Seite `http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm`
- Corpora mailing list
- ACL Corpora Wiki `http://www.aclweb.org/aclwiki/index.php?title=List_of_resources_by_language`
- AMALGAM:
  `http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgover.html`

Zusammenfassung

- Kompetenz und Performanz
- Stärken und Schwächen regelbasierter Systeme
- Stärken und Schwächen statistischer Systeme
- Unterschiede im Ansatz
- Parallel mit Theorien über Spracherwerb
- Korpora und Sprachressourcen