Motivation Statistischer Ansätze und Korpora

Dr. Vera Demberg

Universität des Saarlandes

April 23rd, 2012

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)

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
Prosodie (ToBI)	London-Lund Corpus
Spoken Language	Christine (200,000wds)

(Achtung: keine vollständige Liste! nur Beispiele!)

Entwicklung von Korpora

Annotation: PoS-Tagging

- 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

- sprachspezifisch
- Manche TAGs spiegeln nur Oberflächenform wieder aber desambiguieren nicht:
  - 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

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>.
```

Überblick über Sprachressourcen

Syntax-Korpora ("Baumbanken")

- Penn Treebank: 1M Worte aus dem Wall Street Journal

```
( (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) )
```

- Deutsch:
  - NEGRA
    (20.000 Sätze Frankfurter Rundschau, 400K Worte)
  - TIGER
    (50.000 Sätze Frankfurter Rundschau = 1M Worte)

Überblick über Sprachressourcen

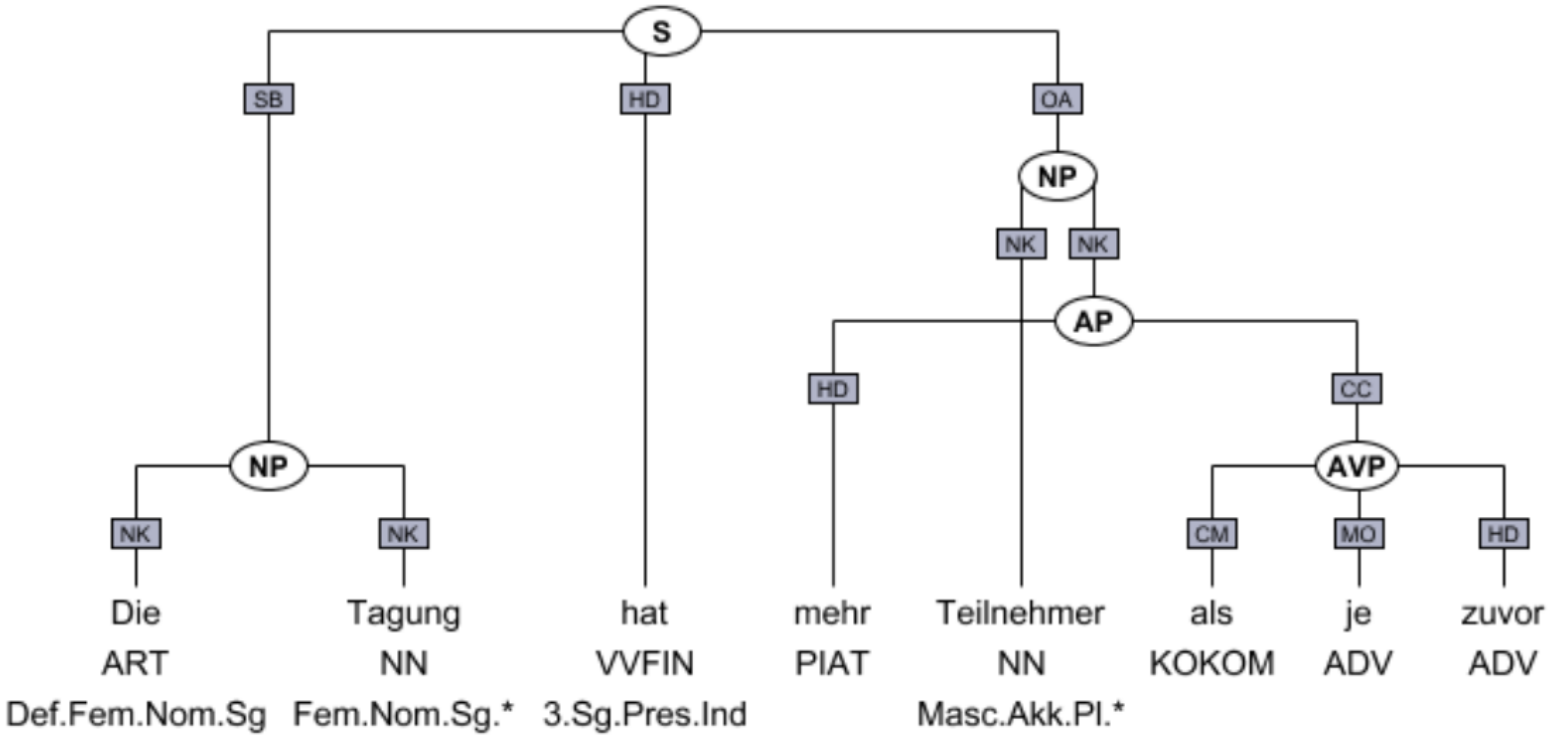
Wall Street Journal

- syntaktische Bäume
- Phänomene:
  - syntaktische Funktionen
  - Nullkonstituenten
  - Spuren, usw.
- Nur finanzielle Texte
- Konvertiert worden zu anderen Formaten (CCG, TAG, HPSG, Dependencies...)
- weitere Annotationsebenen
  (POS-tags, Syntax, Semantische Rollen, Diskurs)

Überblick über Sprachressourcen

Syntax-Korpora ("Baumbanken")

- TIGER
  (50.000 Sätze Frankfurter Rundschau = 1M Worte)



Überblick über Sprachressourcen

Syntax-Korpora ("Baumbanken")

- TIGER
  (50.000 Sätze Frankfurter Rundschau = 1M Worte)

```
<body>

<s id="s5">
  <graph root="s5_504">
    <terminals>
      <t id="s5_1" word="Die" pos="ART" morph="Def.Fem.Nom.Sg"/>
      <t id="s5_2" word="Tagung" pos="NN" morph="Fem.Nom.Sg.*"/>
      <t id="s5_3" word="hat" pos="VVFIN" morph="3.Sg.Pres.Ind"/>
      <t id="s5_4" word="mehr" pos="PIAT" morph="--"/>
      <t id="s5_5" word="Teilnehmer" pos="NN" morph="Masc.Akk.Pl.*"/>
      <t id="s5_6" word="als" pos="KOKOM" morph="--"/>
      <t id="s5_7" word="je" pos="ADV" morph="--"/>
      <t id="s5_8" word="zuvor" pos="ADV" morph="--"/>
    </terminals>
    <nonterminals>
      <nt id="s5_500" cat="NP">
        <edge label="NK" idref="s5_1"/>
        <edge label="NK" idref="s5_2"/>
```

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

```
wsj/00/wsj_0000.mrg 13 8  gold decide.01 —— 7:1-ARG0  8:0-rel          10:1-ARG1
wsj/00/wsj_0000.mrg 13 12 gold slide.01  —— 11:1-ARG0 12:0-rel
wsj/00/wsj_0000.mrg 14 2  gold take.01   —— 1:1-ARG0  2:0-rel          8:2-ARG1
wsj/00/wsj_0000.mrg 14 10 gold get.01    —— 10:0-rel   13:1-ARG2
wsj/00/wsj_0000.mrg 15 2  gold write.01  —— 1:1-ARG0  2:0-rel          9:2-ARG1  10:1-ARGM-LOC
wsj/00/wsj_0000.mrg 16 2  gold eat.01    —— 1:1-ARG0  2:0-rel          9:2-ARG1  11:3-ARGM-CAU
wsj/00/wsj_0000.mrg 16 11 gold get.01    —— 11:0-rel   12:1-ARGM-DIR
wsj/00/wsj_0000.mrg 17 4  gold used.01   —— 3:4-ARG0   4:0-rel          10:2-ARG1 13:2-ARGM-as
wsj/00/wsj_0000.mrg 18 2  gold steal.01  —— 1:1-ARG0  2:0-rel          8:2-ARG1  11:2-ARGM-TMP
```

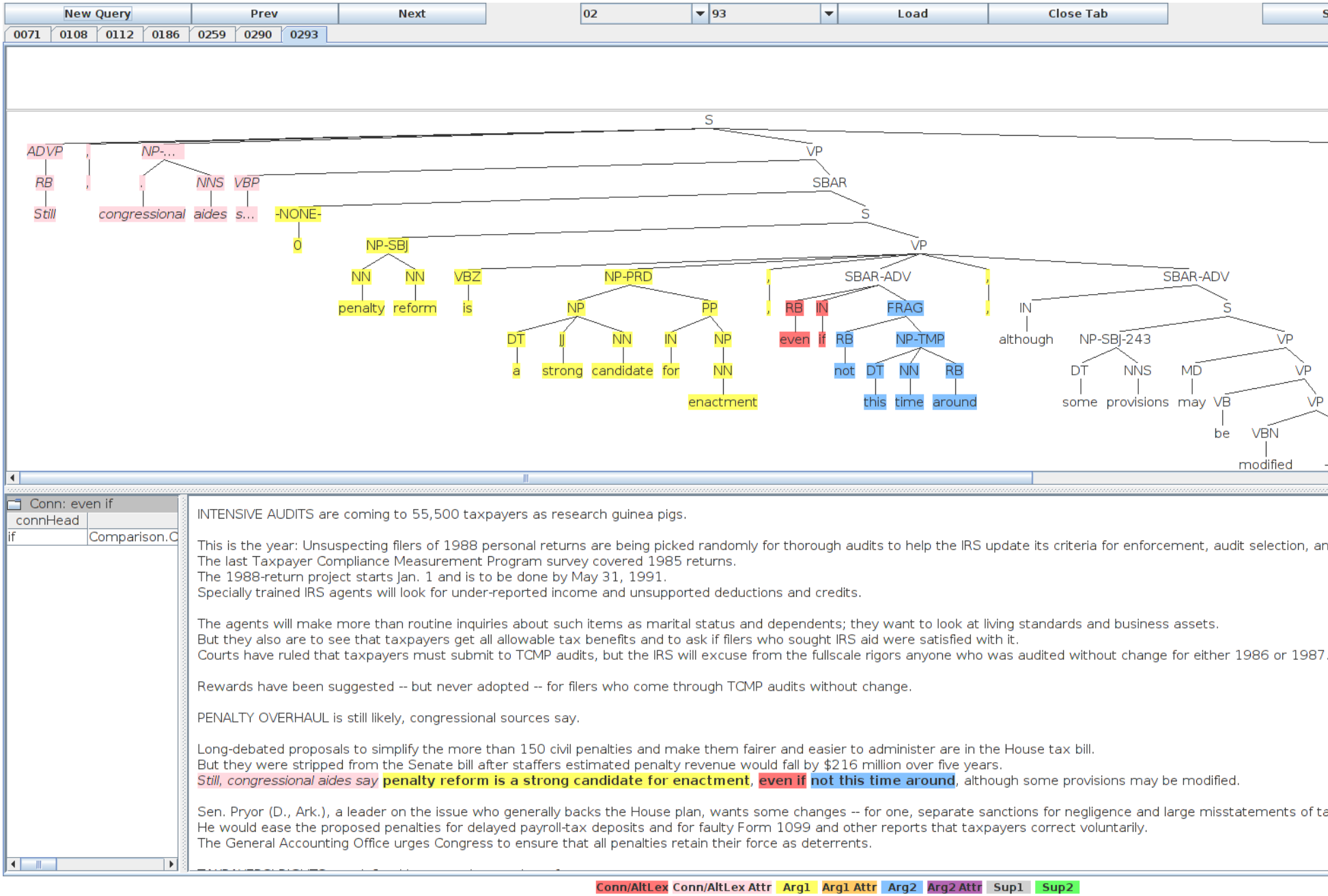
  - Deutsch: SALSA, auf Grundlage von TIGER

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:
  Penn Discourse Bank, auf Grundlage der Penn Treebank

Browser.png


New Query
Prev
Next
02
93
Load
Close Tab
0071
0108
0112
0186
0259
0290
0293
S
ADVP
,
NP-...
VP
RB
,
,
NNS
VBP
SBAR
Still
congressional
aides
s...
-NONE-
S
0
NP-SBJ
VP
NN
NN
VBZ
NP-PRD
,
SBAR-ADV
,
SBAR-ADV
penalty
reform
is
NP
PP
,
RB
IN
FRAG
IN
S
DT
JJ
NN
IN
NP
even
if
RB
NP-TMP
although
NP-SBJ-243
VP
a
strong
candidate
for
NN
not
DT
NN
RB
DT
NNS
MD
VP
enactment
this
time
around
some
provisions
may
VB
VP
be
VBN
modified
Conn: even if
connHead
if
Comparison.C
INTENSIVE AUDITS are coming to 55,500 taxpayers as research guinea pigs.
This is the year: Unsuspecting filers of 1988 personal returns are being picked randomly for thorough audits to help the IRS update its criteria for enforcement, audit selection, an
The last Taxpayer Compliance Measurement Program survey covered 1985 returns.
The 1988-return project starts Jan. 1 and is to be done by May 31, 1991.
Specially trained IRS agents will look for under-reported income and unsupported deductions and credits.
The agents will make more than routine inquiries about such items as marital status and dependents; they want to look at living standards and business assets.
But they also are to see that taxpayers get all allowable tax benefits and to ask if filers who sought IRS aid were satisfied with it.
Courts have ruled that taxpayers must submit to TCMP audits, but the IRS will excuse from the fullscale rigors anyone who was audited without change for either 1986 or 1987.
Rewards have been suggested -- but never adopted -- for filers who come through TCMP audits without change.
PENALTY OVERHAUL is still likely, congressional sources say.
Long-debated proposals to simplify the more than 150 civil penalties and make them fairer and easier to administer are in the House tax bill.
But they were stripped from the Senate bill after staffers estimated penalty revenue would fall by $216 million over five years.
Still, congressional aides say penalty reform is a strong candidate for enactment, even if not this time around, although some provisions may be modified.
Sen. Pryor (D., Ark.), a leader on the issue who generally backs the House plan, wants some changes -- for one, separate sanctions for negligence and large misstatements of ta
He would ease the proposed penalties for delayed payroll-tax deposits and for faulty Form 1099 and other reports that taxpayers correct voluntarily.
The General Accounting Office urges Congress to ensure that all penalties retain their force as deterrents.
Conn/AltLex
Conn/AltLex Attr
Arg1
Arg1 Attr
Arg2
Arg2 Attr
Sup1
Sup2

Korpora mit gesprochener Sprache

London-Lund Corpus: Prosodic annotation

```
1 8 14 1470 1 1 A 11 ^what a_bout a cigar\ette# .                    /
1 8 15 1480 1 1 A 20 *((4 sylls))*                                   /
1 8 14 1490 1 1 B 11 *I ^w\on't have one th/anks#* - - -             /
1 8 14 1500 1 1 A 11 ^aren't you .going to sit d/own# -              /
1 8 14 1510 1 1 B 11 ^[/\m]# -                                       /
1 8 14 1520 1 1 A 11 ^have my _coffee in p=eace# - - -                /
1 8 14 1530 1 1 B 11 ^quite a nice .room to !s\it in ((actually))#   /
1 8 14 1540 1 1 B 11 *^\isn't* it#                                   /
1 5 15 1550 1 1 A 11 *^y/\es#* - - -                                 /

The codes used in this example are:

# end of tone group
^ onset
/ rising nuclear tone
\ falling nuclear tone
/\ rise-fall nuclear tone
_ level nuclear tone
[] enclose partial words and phonetic symbols
. normal stress
! booster: higher pitch than preceding prominent syllable
= booster: continuance
(( )) unclear
* * simultaneous speech
- pause of one stress unit
```

Korpora mit gesprochener Sprache

London-Lund Corpus: Prosodic annotation

```
1 8 14 1470 1 1 A 11 ^what a bout a cigar\ette# .                    /
1 8 15 1480 1 1 A 20 *((4 sylls))*                                   /
1 8 14 1490 1 1 B 11 *I ^w\on't have one th/anks#* - - -             /
1 8 14 1500 1 1 A 11 ^aren't you .going to sit d/own# -              /
1 8 14 1510 1 1 B 11 ^[/\m]# -                                       /
1 8 14 1520 1 1 A 11 ^have my _coffee in p=eace# - - -                /
1 8 14 1530 1 1 B 11 ^quite a nice .room to !s\it in ((actually))# /
1 8 14 1540 1 1 B 11 *^\isn't* it#                                   /
1 5 15 1550 1 1 A 11 *^y/\es#* - - -                                 /

The codes used in this example are:

# end of tone group
^ onset
/ rising nuclear tone
\ falling nuclear tone
/\ rise-fall nuclear tone
_ level nuclear tone
[] enclose partial words and phonetic symbols
. normal stress
! booster: higher pitch than preceding prominent syllable
= booster: continuance
(( )) unclear
* * simultaneous speech
- pause of one stress unit
```

Spezielle Schwierigkeiten bei der Annotation prosodischer Korpora:

- Beurteilung beruht auf subjektivem Eindruck (z.B. Realisierung von Akzenten und Betonungen)
- Wenn im Wort annotiert wird, schwierig Annotation und Original wieder auseinanderzudividieren.
- Sonderzeichen können zu Schwierigkeiten in der Darstellung und Verarbeitung führen.

Zusammenfassung linguistische Annotation

- verschiedene Level von linguistischer Annotation
- Entscheidungen bei Annotation beeinflusst spätere Anwendungen
- Annotationsguidelines
  - Annotationsaufgabe so formulieren, dass es hohes Interannotator agreement gibt.
- Formatierung
- Veröffentlichung (z.B. über LDC)

Praktisch gesehen...

Wie können wir Text-Korpora nützen?

- Ausbeuten reinen Texts
  - Durch Unix-Tools: Besonders effizient durch C-Implementation und die Möglichkeit, Befehlsketten zu bauen
  - Mit geeigneten Programmiersprachen, etwa sed, awk, perl, python, java, usw.
- Aufbereitung von Texten (durch Tagger, Chunker, Parser, usw.)
- Auslesen hinzugefügter linguistischen Information
  (mit spezifischen Tools, z. B. Konkordanz-Programmen, oder mit allgemeinen Programmiersprachen)

Word Counts

- Word Tokens: **Gesamtzahl** Wörter im Korpus

  $Peters_1$ $Vater_2$ ist_3 $Koch_4$ $.5$ $Peters_6$ $Mutter_7$ ist_8 $Köchin_9$ $.10$

- Word Types: Anzahl **verschiedener** Wörter im Korpus

  $Peters_1$ $Vater_2$ ist_3 $Koch_4$ $.5$ $Peters_1$ $Mutter_6$ ist_3 $Köchin_7$ $.5$

  Hier muss aber entschieden werden, was als gleich zählt:

  - *Peter* und *Peters*? *bin, bist, ist,...*? *Koch* und *Köchin*? *Vater* und *Mutter*?
  - Homographen verschiedener Wortkategorien?
    $to\ saw_V$ *the wood/sharpen the* saw_N
    das $schnelle_{ADJ}$ *Auto/der Zug fährt* $schnell_{ADV}$
    (Voraussetzung, das Korpus ist getaggt)
  - Homographen mit verschiedener Bedeutung?
    saw the wood/saw the film

- Das Type/Token Ratio: Kann zur Charakterisierung von Texten, Genres, Autoren, usw. dienen

Erfolg und Nutzung von Korpora

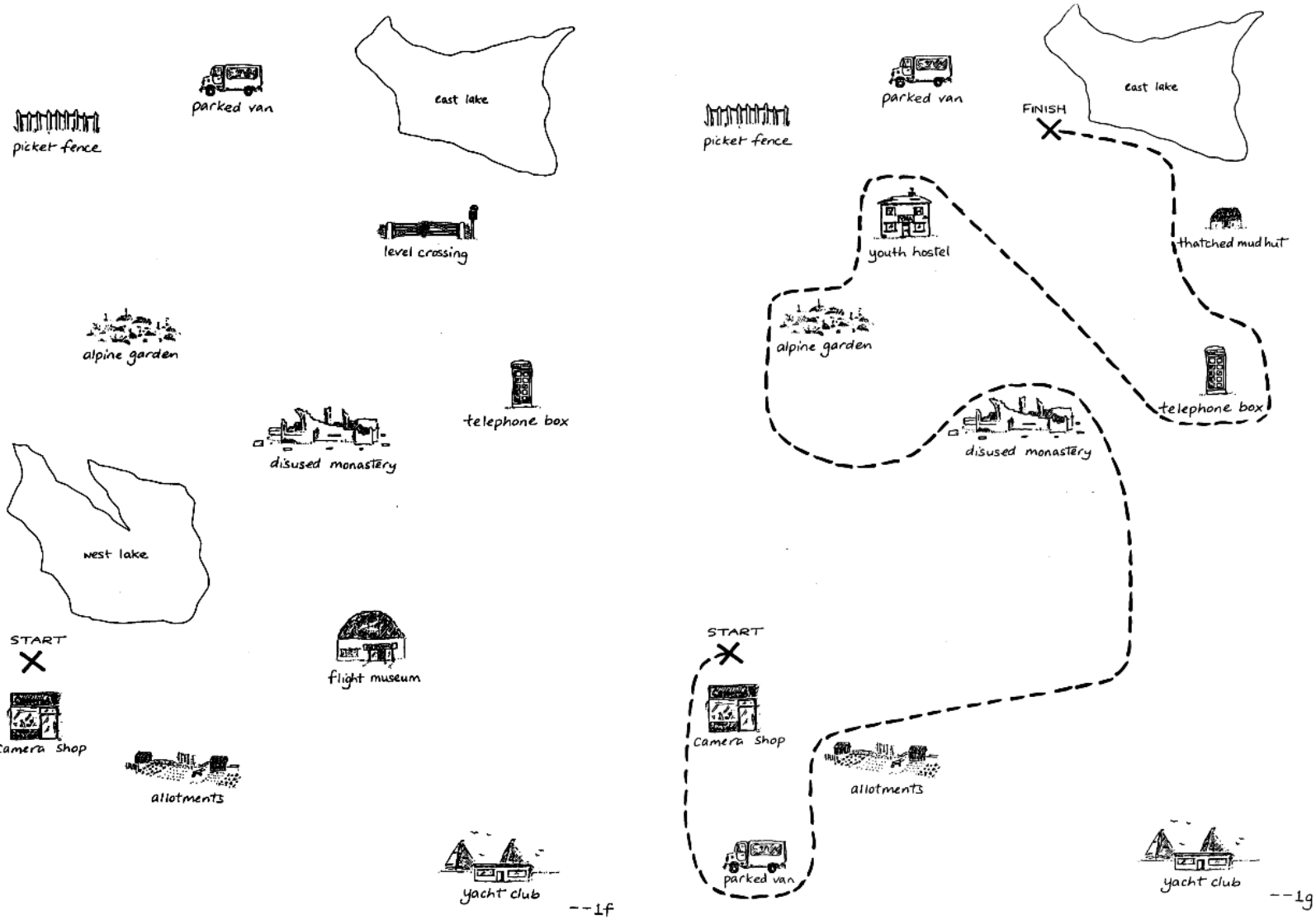
- Suche in Korpora (tgrep2, Tiger search u.v.m)
- automatische Tools entwickeln:
  Penn Treebank hat Parsing revolutioniert und statistische Ansätze erfolgreich gemacht.
- Wichtig für Evaluation: gold standard zum Vergleich verschiedener Ansätze auf bestimmten Daten
- Evaluation von Theorien über Sprache (Beispiele für bestimmte Phänomene finden)

Überblick über Sprachressourcen (2)

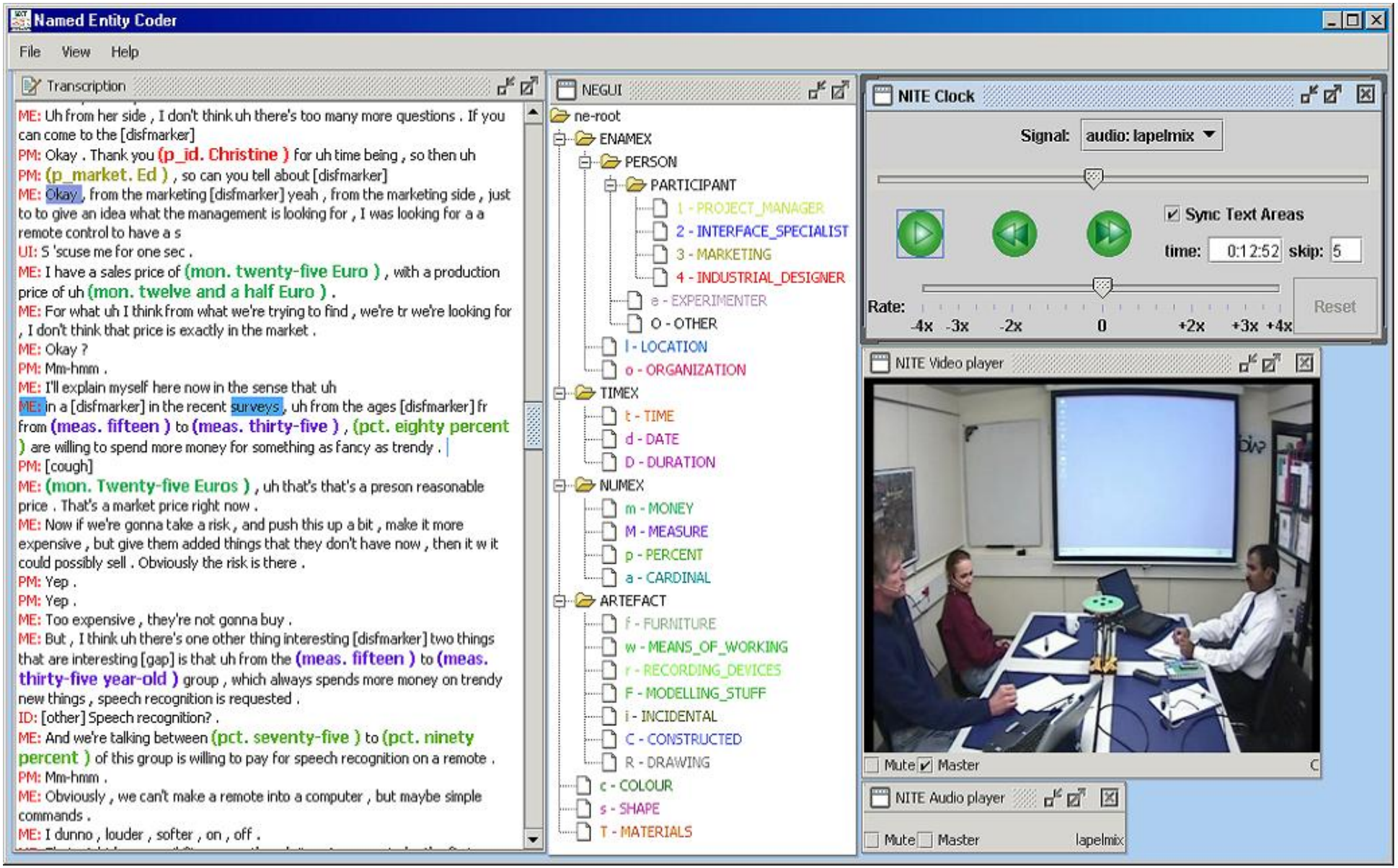
Korpora mit unterschiedlichen Modalitäten:

Annotationsart	Corpus
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)
Child speech	CHILDES (child-directed speech)

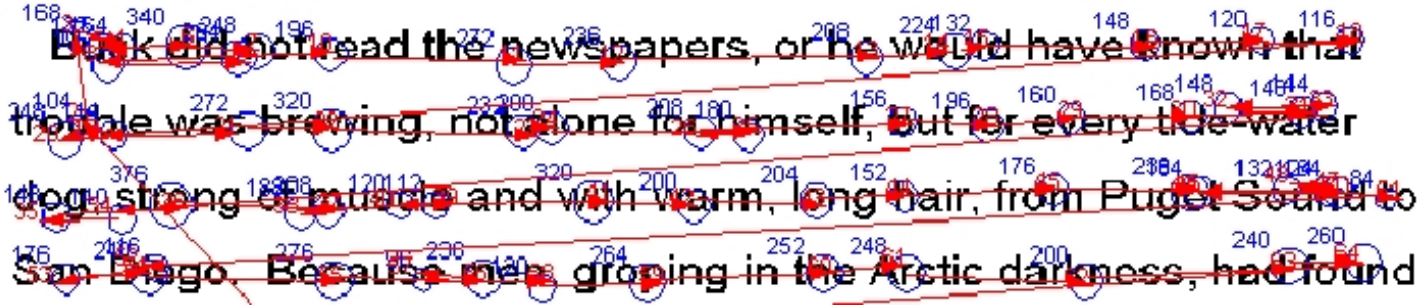
MapTask Corpus



AMI Corpus


Named Entity Coder
File View Help
Transcription
ME: Uh from her side , I don't think uh there's too many more questions . If you can come to the [disfmarker]
PM: Okay . Thank you (p_id. Christine) for uh time being , so then uh
PM: (p_market. Ed) , so can you tell about [disfmarker]
ME: Okay , from the marketing [disfmarker] yeah , from the marketing side , just to to give an idea what the management is looking for , I was looking for a a remote control to have a s
UI: S 'scuse me for one sec .
ME: I have a sales price of (mon. twenty-five Euro) , with a production price of uh (mon. twelve and a half Euro) .
ME: For what uh I think from what we're trying to find , we're tr we're looking for , I don't think that price is exactly in the market .
ME: Okay ?
PM: Mm-hmm .
ME: I'll explain myself here now in the sense that uh
ME: in a [disfmarker] in the recent surveys , uh from the ages [disfmarker] fr from (meas. fifteen) to (meas. thirty-five) , (pct. eighty percent) are willing to spend more money for something as fancy as trendy .
PM: [cough]
ME: (mon. Twenty-five Euros) , uh that's that's a preson reasonable price . That's a market price right now .
ME: Now if we're gonna take a risk , and push this up a bit , make it more expensive , but give them added things that they don't have now , then it w it could possibly sell . Obviously the risk is there .
PM: Yep .
PM: Yep .
ME: Too expensive , they're not gonna buy .
ME: But , I think uh there's one other thing interesting [disfmarker] two things that are interesting [gap] is that uh from the (meas. fifteen) to (meas. thirty-five year-old) group , which always spends more money on trendy new things , speech recognition is requested .
ID: [other] Speech recognition? .
ME: And we're talking between (pct. seventy-five) to (pct. ninety percent) of this group is willing to pay for speech recognition on a remote .
PM: Mm-hmm .
ME: Obviously , we can't make a remote into a computer , but maybe simple commands .
ME: I dunno , louder , softer , on , off .
NEGUI
ne-root
ENAMEX
PERSON
PARTICIPANT
1 - PROJECT_MANAGER
2 - INTERFACE_SPECIALIST
3 - MARKETING
4 - INDUSTRIAL_DESIGNER
e - EXPERIMENTER
O - OTHER
l - LOCATION
o - ORGANIZATION
TIMEX
t - TIME
d - DATE
D - DURATION
NUMEX
m - MONEY
M - MEASURE
p - PERCENT
a - CARDINAL
ARTEFACT
f - FURNITURE
w - MEANS_OF_WORKING
r - RECORDING_DEVICES
F - MODELLING_STUFF
i - INCIDENTAL
C - CONSTRUCTED
R - DRAWING
c - COLOUR
s - SHAPE
T - MATERIALS
NITE Clock
Signal: audio: lapelmix
Sync Text Areas
time: 0:12:52 skip: 5
Rate:
-4x -3x -2x 0 +2x +3x +4x
Reset
NITE Video player
Mute Master
NITE Audio player
Mute Master lapelmix

Dundee Eye-tracking Corpus



WORD	TEXT	LINE	OLEN	WLEN	XPOS	WNUM	FDUR	OBLP	WDLP	LAUN	TXFR
*Blink	1	-99	0	0	-99	-99	92	-99	-99	-99	-99
Rather	1	1	6	6	2	1	165	2	2	-99	42
Rather	1	1	6	6	2	1	58	2	2	0	42
Rather	1	1	6	6	2	1	154	2	2	0	42
as	1	1	2	2	9	2	271	2	2	-7	422
Basil	1	1	5	5	12	3	257	2	2	-3	1
Fawlty	1	1	6	6	19	4	401	3	3	-7	1
couldn't	1	1	8	8	26	5	291	3	3	-7	7
stop	1	1	4	4	36	6	200	4	4	-10	10
stop	1	1	4	4	34	6	150	2	2	2	10
talking	1	1	7	7	40	7	211	3	3	-6	5
about	1	1	5	5	47	8	236	2	2	-7	164
war	1	1	3	3	55	10	344	0	0	-8	28
his	1	1	3	3	62	12	192	0	0	-7	161
German	1	1	6	6	70	13	263	4	4	-8	7
guests,	1	1	7	6	78	14	176	5	5	-8	3

CHILDES Corpus

```
*ELS:   are you gonna [: going to] do this one ?
%xmor:  v:aux|be&PRES pro|you part|go-PROG inf|to v|do det|this pro:indef|one ?
*MOT:   yes .
%xmor:  co|yes .
*MOT:   you can see if you want .
%xmor:  pro|you v:aux|can v|see conj:subor|if pro|you v|want .
*CHI:   see there .
%xmor:  v|see adv:loc|there .
*CHI:   see there . [+ SR]
%xmor:  v|see adv:loc|there .
*ELS:   see there ?
%xmor:  v|see adv:loc|there ?
*CHI:   yes .
%xmor:  co|yes .
*CHI:   all fall down see .
%xmor:  qn|all v|fall adv|down co|see .
*ELS:   oh .
%xmor:  co|oh .
*ELS:   all fall down see .
%xmor:  qn|all v|fall adv|down co|see .
*ELS:   look .
%xmor:  v|look .
*ELS:   have a look under the table .
%xmor:  v|have det|a n|look prep|under det|the n|table .
*ELS:   all fall down .
```

Korpus Vorbereitung

Für manche Analysen muss ein Korpus aufbereitet werden.

Typische Aufbereitungsschritte:

- Satzgrenzen bestimmen
  I spoke to Mr. and Mrs. Gore from Washington D.C. today.
  - Baseline: Satzgrenze nach jedem Punkt, Fragezeichen, Ausrufezeichen: korrekt in 90% der Fälle.
  - Regeln mit Liste von typischen Abkürzungen und Heuristik ueber kleingeschriebene Wörter nach Satzzeichen.
  - Statistische Ansätze am besten (bis zu 99.25% korrekt).
- Wortgrenzen bestimmen
- eventuell Säubern
  - *hmm* in gesprochener Sprache
  - überlappende Äusserungen
  - Titel, Inhaltsverzeichnisse, XML, Fussnoten, Tabellen
- Lemmatisierung

Korpus-Aufbereitung: Tokenisierung

Frage: Was ist ein Wort?
Eine Folge alphanumerischer Zeichen, durch Leerzeichen (allg.) oder Interpunktion getrennt

- Interpunktion von Wörtern trennen, wenn keine Abkürzung
- Apostrophen ersetzen? *Robert's*, *isn't* vs. *it's*, *geht's*
- Striche interpretieren: Gedankenstrich, Komposita, Bindestrich, Umbrüche an Zeilenenden
  co-operate, e-mail, text-based, so-called, pro-Arab
- Spezielle Tokens: Zahlen (2 000), "Named entities" (Namen, Daten, Zitate, Adressen, Telefonnummer, usw.), multi-word expressions

Annotation: Lemmatisierung

- Def.: Wörter mit deren Grundform versehen
- Probleme bei Homographen, s. oben
- Vereinfachungsmöglichkeiten:
  - Grossschreibung eliminieren
  - Satzinterne Interpunktion löschen
  - Stemming (Affixe löschen) z.B. *operating, operate, operated, operates, operator: operat*
- Wird heutzutage selten gemacht, weil zu viele wichtige Information verloren geht.

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

- Korpora sind extrem wichtig für Computerlinguisten (Entwicklung von Tools, Evaluation, Validierung von Theorien, Sichtung von Daten)
- Wichtige Aspekte zur Nutzung von Korpora
  - Art der Sprache (Dialog, gesprochene vs. geschriebene Sprache, Kindliche Sprache, ...)
  - Ist ein bestimmtes Genre wichtig? Ausgewogenes Korpus?
  - Ist das Korpus gross genug?
  - Hat es die richtige Annotation, oder muss es (automatisch) vorverarbeitet werden?
  - Welche Art von Statistik soll extrahiert werden?
  - Welche Tools können dabei helfen?
  - Haben die existierenden Tools das erwartete Verhalten (Wie wird ein Wort, Satz etc. gezählt)?