Mathematische Grundlagen III

Korpora und Sprachressourcen

Garance PARIS

26. April 2011

Sprachressourcen

Überblick

- Korpora (Singular Korpus, neutrum!):
  Große Mengen von Text
- 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 Kučera: Ersteller des Brown Corpus, eins der ersten Korpora fürs Englische)

Korpora

Definition

- Die Daten sollten aus möglichst natürlichen Gesprächssituationen stammen
- Meist ist es wünschenswert, sich auf Muttersprachler zu begrenzen
- Sie sollten repräsentativ sein, wobei sich die Frage stellt, was denn repräsentativ ist!
- Heutzutag sind Korpora nur noch elektronisch vorstellbar
- Sie werden manchmal für einen bestimmten Zweck erhoben (z. B. Fehleranalyse bei Sprachlernern im CALL-Bereich)

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 Kurzfilm . 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 Verpflichtung 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)
- 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

Aufbereitung: Annotation

Annotation: Hinzufügen linguistischer Information

Probleme:

- Welcher Tagset, welcher (Grammatik-) Formalismus, ...?
- Interpretation (HPSG/LFG vs. funktionale Grammatik)
- Wegen Ambiguität ist Annotation nicht einfach
- Es ist aufwendig, da Handarbeit
  - Annotationsaufwand für ein Wort: 30 Sekunden
  - 1M Worte: 500 000 Minuten = 5 Jahre
  - Plus Aufwand fuer Qualitätssicherung
- Außerdem ist es auch fehlerbehaftet und möglicherweise inkonsistent
- Aber automatische Annotation ist nicht 100 % zuverlässig und macht systematische Fehler

Mögliche Lösungen

- Annotationsmöglichkeiten gering halten
  (z. B. kleines Tagset), um schwierige Entscheidungen aus dem Weg zu gehen
- Bei Unsicherheit mehrere Tags zuweisen
  (dem User wissen lassen, 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

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 (s. oben), Text oder Binär (indexiert)
- Medium (Text, Audio, Transkripte, Video, usw.)
- Aufbereitung und Annotation
- Urheber- und Nutzungsrechte, Preis
- Standard-Referenz: Allgemeine Verfügbarkeit

Überblick über Sprachressourcen

Korpora mit Wortarten

- Standardkorpora:
  - British National Corpus (BNC), 100M Worte
  - American National Corpus (ANC), 22M Worte
  - Huge German Corpus (HGC), 200M Worte
- Einsatz: Training von Taggern

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