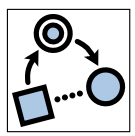
Komplexität und natürliche Sprache

Komplexität natürlichsprachlicher Sätze

Timm Lichte & Christian Wurm

Heinrich-Heine-Universität Düsseldorf

Sommersemester 2017, 21.06.2017



Outline

Outline

Rekapitulation: Quiz

$L_1 = ba^*$
$L_2 = ba^{\leq 2000}$

algorithmische Komplexität:	L_1	???	L_2
Kolmogorov-Komplexität:	L_1	???	L_2
Entropie:	L_1	???	L_2

Anwendung auf natürliche Sprache

Konkrete Anwendungsfälle mit Bezug auf natürliche Sprache

- Klasse der natürlichen Sprachen: schwach kontextsensitiv, RTTM-erkennbar, … (Sitzung am 03.05.)

Heute

- Vergleich natürlichsprachlicher Worte (=Sätze, Phrasen)
  - Zum Beispiel: berechne deren Gesamtinformativität
  - Problem: Phrasenstruktur, Satzbedeutung

Nächste Woche

- Vergleich natürlicher Sprachen (Sitzung nächste Woche)
  - Zum Beispiel: bilde die kürzesten TM's und vergleiche deren Länge
  - Problem: Durchführung, kulturtheoretische Implikationen

Outline

Komplexität von Sätzen

Was macht einen Satz intuitiv komplex:

- Satzlänge
- Verschachtelung (z.B. Zentraleinbettung)
- ungewöhnliche Worte und Konstruktionen
- hohe Ambiguität (syntaktisch und semantisch)
- semantische Brüche, Inkohärenz
- ungewöhnliche Assoziationen

Schwarze Milch der Frühe wir trinken sie abends
wir trinken sie mittags und morgens wir trinken sie nachts
wir trinken und trinken
wir schaufeln ein Grab in den Lüften da liegt man nicht eng
...

Komplexität von Sätzen

Was macht einen Satz intuitiv komplex:

- Satzlänge
- Verschachtelung (z.B. Zentraleinbettung)
- ungewöhnliche Worte und Konstruktionen
- hohe Ambiguität (syntaktisch und semantisch)
- semantische Brüche, Inkohärenz
- ungewöhnliche Assoziationen

Schwarze Milch der Frühe wir trinken sie abends
wir trinken sie mittags und morgens wir trinken sie nachts
wir trinken und trinken
wir schaufeln ein Grab in den Lüften da liegt man nicht eng
...

(Paul Celan, Die Todesfuge)

Algorithmische Komplexität von Sätzen

(1) a. The first shot the tired soldier the mosquito bit fired missed.
b. The first shot fired by the tired soldier bitten by the mosquito missed.

Komplexität = algorithmische Komplexität

Gegeben eine Turingmaschine M, die algorithmische Komplexität eines Satzes $\overline{w}$ ist der Zeit- und Speicherverbrauch von M bei der Erkennung von $\overline{w}$.

Probleme:

- Was ist M? Soll M die ganze natürliche Sprache von $\overline{w}$ erkennen?
- Was ist Σ? "wort"- oder "buchstaben"-basiert?
- Wie sollen Zeit- und Speicherverbrauch miteinander verrechnet werden?

Informativität von Sätzen

(2) a. Das Kamel stieß das Zebra weg.
b. Das Zebra wurde vom Kamel weggestoßen.
c. Es wurde von ihm weggestoßen.
d. Von ihm.

Komplexität = Summe der Informativität der einzelnen Buchstaben

- $\Sigma = \{A, \ldots, z, ., \text{␣}\}$
- $I_{\overline{w}} = \sum_{i=1}^{n} I(a_i)$ mit $I(a_i) = -\log P(a_i)$
- $P_{\overline{w}}(\sigma) = \frac{|\overline{w}|_\sigma}{|\overline{w}|}$

Oder:

- $\hat{P}_L(\sigma) = lim_{n\to\infty} \sum_{\overline{w}\in L_n} \frac{|\overline{w}|_\sigma}{|\overline{w}|}$
- Man kann $\hat{P}_L$ mit einem großen Sprachkorpus schätzen.

Informativität von Sätzen

Probleme der Informativität:

- Das Auftreten eines Buchstabens ist entscheidend, aber nicht die Struktur der Sätze.
- Mit $P_{\overline{w}}(\sigma)$ ist Informativität stark abhängig vom Wort.
- $\hat{P}_L(\sigma)$ ohne große Sprachkorpora nicht schätzbar.
- "Worte" als Buchstaben? $\Sigma = \{\text{Das}, \text{Kamel}, \text{stieß}, \ldots\}$
- Was bedeutet es für einen Satz, eine bestimmte Informativität zu haben?
- Semantik spielt gar keine Rolle.

Linguistische/Angewandte Komplexitätsmaße für Sätze

Aus verschiedenen Gründen (v.a. Durchführbarkeit und Datenfokus) konzentrieren sich linguistische/angewandte Komplexitätsmaße im Allgmeinen auf bestimmte Teilaspekte von Sprache und Sätzen.

Beispiele:

- quantitative Lesbarkeitsmetriken
- (phrasen-)strukturelle Metriken
  - Dependency Locality Theory / Syntactic Prediction Locality Theory (Gibson [2; 3])
  - Phrasal Combination Domain, IC-to-word ratios, missassignment (Hawkins [4])
  - Surprisal (z.B. Demberg & Keller [1])

Outline

Quantitative Lesbarkeitsmetriken

- populär im angel-sächsischen Raum (mindestens seit den 40ern)
- Einsatz im Militär, Gesundheitswesen und Bildungswesen zur Bewertung von Texten und Lesefähigkeiten
- Teil von Textverarbeitungsprogrammen wie MS Word

Beispiele:

- Flesch-Kincaid readability tests
- Dale-Chall readability formula
- Gunning fog index
- McLaughlin's SMOG formula
- Lexile score
- CohMetrix
- ...

Flesch-Kincaid readability tests

Entwickelt 1975 in der US Navy, aufbauend auf dem Flesch-Test.

Flesch–Kincaid grade level

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

(3) The Australian platypus is seemingly a hybrid of a mammal and reptilian creature.

= 11.3 (24 Silben, 13 Worte)

School Level	Notes
5th grade	Very easy to read. Easily understood by an average 11-year
6th grade	Easy to read. Conversational English for consumers.
7th grade	Fairly easy to read.
8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old stud
10th to 12th grade	Fairly difficult to read.
College	Difficult to read.
College Graduate	Very difficult to read. Best understood by university gradua

Dale-Chall readability score

Entwickelt in den 40ern, aufbauend auf dem Flesch-Test.

Dale-Chall readability score (für Texte)

1. Select several 100-word samples throughout the text.
2. Compute the average sentence length in words (divide the number of words by the number of sentences).
3. Compute the percentage of words NOT on the Dale–Chall word list of 3,000 easy words.
4. Compute this equation
$$0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

(4) The Australian platypus is seemingly a hybrid of a mammal and reptilian creature.

= 8.509 (13 Worte, 7 schwere Worte)

Dale-Chall readability score (cont.)

Score	Notes
4.9 or lower	easily understood by an average 4th-grade student or lower
5.0–5.9	easily understood by an average 5th or 6th-grade student
6.0–6.9	easily understood by an average 7th or 8th-grade student
7.0–7.9	easily understood by an average 9th or 10th-grade student
8.0–8.9	easily understood by an average 11th or 12th-grade student
9.0–9.9	easily understood by an average 13th to 15th-grade (college) stude

Korrelation mit Lesetests: 0.93

McLaughlin's SMOG formula

- SMOG = Simple Measure of Gobbledygook
- entwickelt 1969 von Harry McLaughlin

SMOG grade (für Texte)

1. Count a number of sentences (at least 30)
2. In those sentences, count the polysyllables (words of 3 or more syllables).
3. Calculate SMOG using

$$1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

(5) The Australian platypus is seemingly a hybrid of a mammal and reptilian creature.

= 14.555 (4 Dreisilber)

McLaughlin's SMOG formula (cont.)

Total Polysyllabic Word Count	Approximate Grade Level
1–6	5
7–12	6
13–20	7
21–30	8
31–42	9
43–56	10
57–72	11
73–90	12
91–110	13
111–132	14
133–156	15
157–182	16
183–210	17
211–240	18

Korrelation mit Lesetests: 0.88

Sonstige

Lexile measure (1988)

The Lexile framework uses average sentence length, and average word frequency in the American Heritage Intermediate Corpus to predict a score on a 0–2000 scale. The AHI Corpus includes five million words from 1,045 published works often read by students in grades three to nine.

CohMetrix (2003)

Standard text readability formulas scale texts on difficulty by relying on word length and sentence length, whereas Coh-Metrix is sensitive to cohesion relations, world knowledge, and language and discourse characteristics.

- Nutzt etwa 200 Maße.

Outline

(Phrasen-)Strukturelle Metriken

Datenfokus: v.a. Relativsatzkonstruktionen (Subjekt, Objekt, reduziert)

(6) a. The reporter [who attacked the senator] admitted the error.
b. The reporter [who the senator attacked] admitted the error.

(7) a. The intern [who the nurse supervised] had bothered the administrator [who lost the medical reports].
b. #The administrator [who the intern [who the nurse supervised] had bothered] lost the medical reports.

(8) The horse raced past the barn fell.

Zuhilfenahme der Phrasenstruktur/Dependenzstruktur eines Satzes.

Dependency Locality Theory (DLT): Integrationskosten

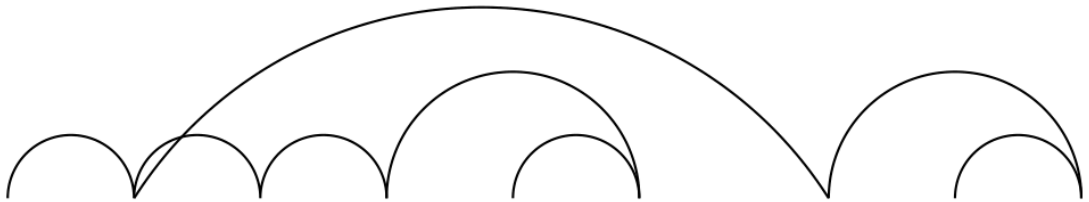
(Gibson 1998; 2000)

Integrationskosten (Integration Cost)

The integration cost associated with integrating a new input head h_2 with a head h_1 that is part of the current structure for the input consists of two parts:
(1) a cost dependent on the complexity of the integration (e.g. constructing a new discourse referent); plus
(2) a **distance-based cost**: a monotone increasing function I(n) energy units (EUs) of the number of new discourse referents that have been processed since h_1 was last highly activated. For simplicity, it is assumed that I(n) = n EUs.

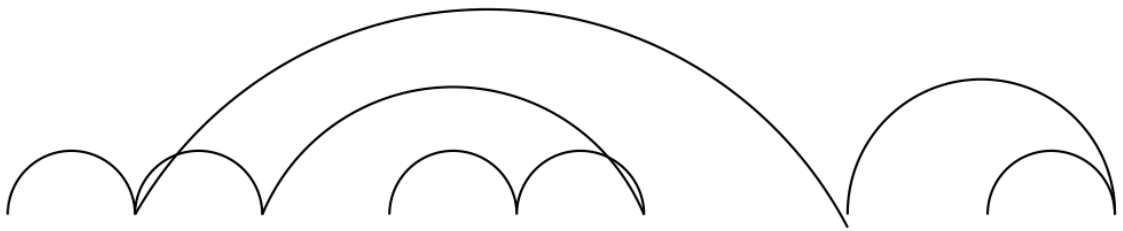
Dependency Locality Theory (DLT): Integrationskosten (cont.)

SRC:



	The	reporter	who	attacked	the	senator	admitted	the	error.
Disc.ref.		x		x		x	x		x
Integ. Cost	I(0)	I(1)	I(0)	I(1)	I(0)	I(1)	I(1)+I(3)	I(0)	I(1)

ORC:



	The	reporter	who	the	senator	attacked	admitted	the	error.
Disc.ref.		x		x		x	x		x
Integ. Cost	I(0)	I(1)	I(0)	I(0)	I(1)	I(1)+I(2)	I(1)+I(3)	I(0)	I(1)

(beide Graphiken aus Demberg & Keller [1])

Hawkins-Maße

(Hawkins 2014)

Fokus: Vorhersage von Linearisierungspreferenzen und Holzwegkonstruktionen

(9) a. The man[$_{VP}$ looked [$_{PP1}$ for his son] [$_{PP2}$ in the dark and quite derelict building]]
b. The man[$_{VP}$ looked [$_{PP2}$ in the dark and quite derelict building] [$_{PP1}$ for his son]]

⇒ IC-to-word ratios

(10) a. I believe the boy knows the answer
b. The horse raced past the barn fell

⇒ Missassignment measures

Hawkins-Maße: IC-to-word ratios

Prasal Combination Domain (PCD)

The PCD for a mother node M and its I(mmediate) C(onstituent)s consists of the smallest string of terminal elements (plus all M-dominated non-terminals over the terminals) on the basis of which the processor can construct M and its ICs.

(11) a. The man [$_{VP}$ looked [$_{PP1}$ for his son] [$_{PP2}$ in the dark and
1 2 3 4 5
quite derelict building]]

b. The man [$_{VP}$ looked [$_{PP2}$ in the dark and quite derelict
1 2 3 4 5 6 7
building] [$_{PP1}$ for his son]]
8 9

Hawkins-Maße: IC-to-word ratios (cont.)

IC-to-word ratio

IC-to-word ratio = n/m with n ICs to be recognized on the basis of m words in the terminal string

Early Immediate Constituents (EIC) [Hawkins 1994: 69–83]

The human processor prefers linear orders that minimize PCDs (by maximizing their IC-to-word ratios), in proportion to the minimization difference between competing orders.

Präferenzen bei PP-Einbettung:

(6)

a. $[_{VP}$ went $[_{PP}$ to the movies$]]$
1 2

b. $[[$the movies to $_{PP}]$ went $_{VP}]$
1 2

c. $[_{VP}$ went $[$the movies to $_{PP}]]$
1 2 3 4

d. $[[_{PP}$ to the movies$]$ went $_{VP}]$
1 2 3 4

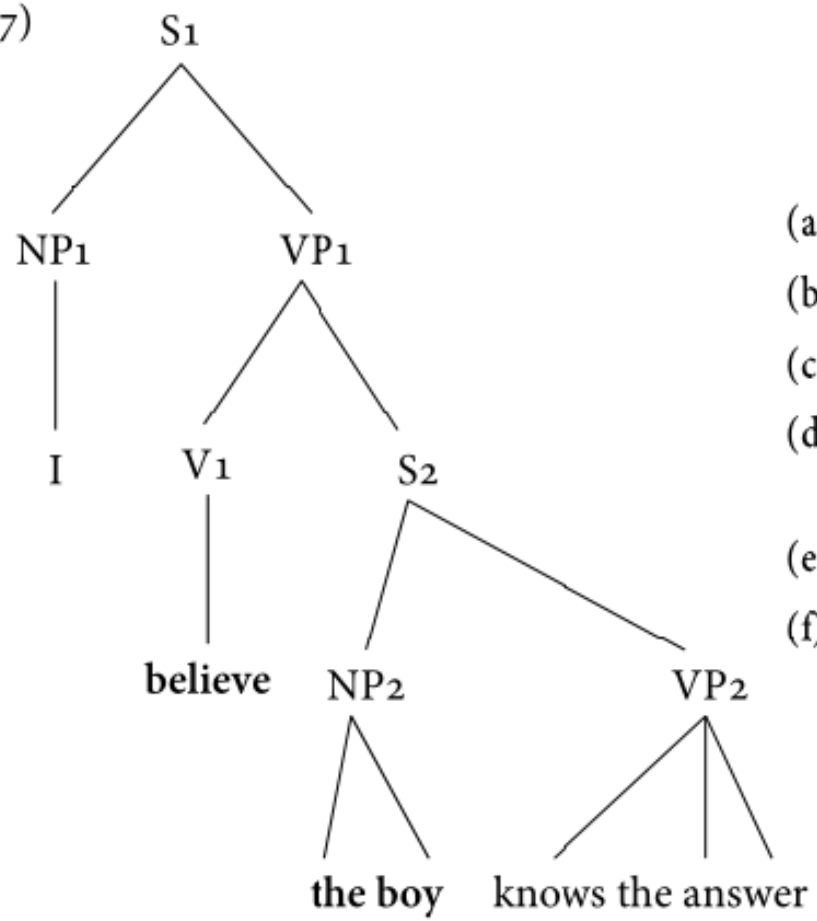
Hawkins-Maße: missassignment measure

Misassignment factors

(a) the number of words and phrases that undergo some temporary misassignment of properties on-line;
(b) the number of any additional dominating nodes that must be introduced into the syntactic tree when correcting the misassignments in (a);
(c) the number of any mother-daughter attachments that are temporarily misassigned to the words and phrases in (a);
(d) the number of any relations of combination or dependency that are temporarily misassigned to the words and phrases in (a);
(e) the number of mother-daughter attachments that replace those misassigned in (c);
(f) the number of relations of combination or dependency that replace those misassigned in (d).

Hawkins-Maße: missassignment measure (cont.)

(27)

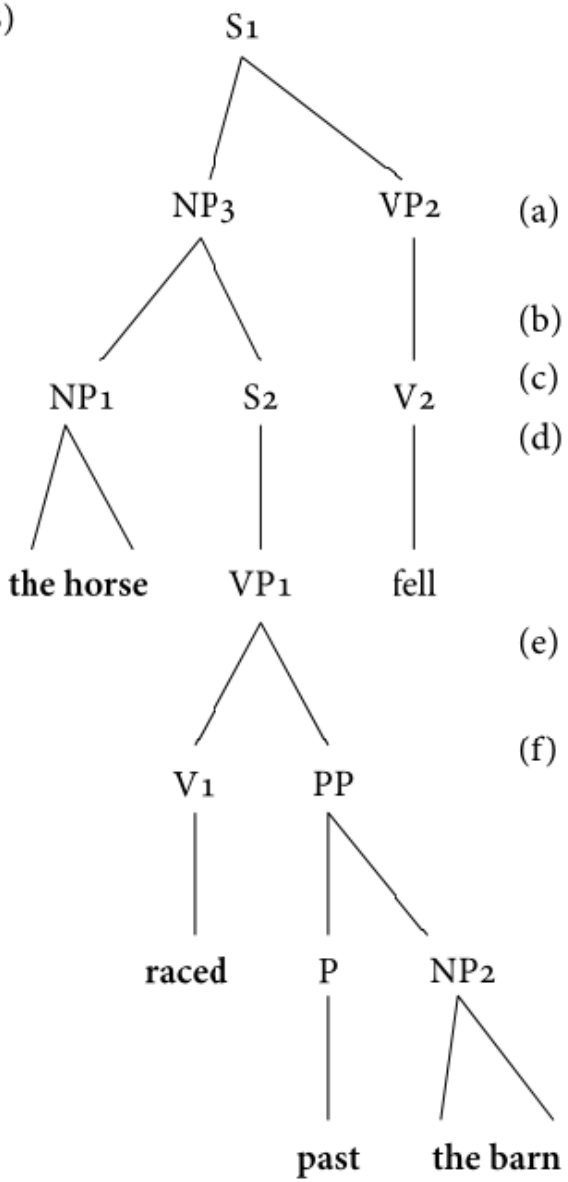


(a)	the, boy, NP2	[3]
(b)	S2	[1]
(c)	$[_{VP1}$ NP2]	[1]
(d)	NP2 = Accus of believe NP2 = θ-object of believe	[2]
(e)	$[_{S2}$ NP2], $[_{VP1}$ S2]	[2]
(f)	NP2 = Nomin of knows NP2 = θ-subject of knows	[2]
	TOTAL = 11	

(aus Hawkins [4])

Hawkins-Maße: missassignment measure (cont.)

(28)



(a)	the, horse, raced, past, the barn, NP1, VP1, PP, NP2	[10]
(b)	NP3, S2	[2]
(c)	[S1 NP1], [S1 VP1]	[2]
(d)	VP1 = predication on NP1 VP1 = active VP NP1 = Nomin of raced NP1 = θ-subject of raced	[4]
(e)	[NP3 NP1], [S1 NP3], [S2 VP1],[NP3 S2]	[4]
(f)	VP1 = rel clause modifying horse VP1 = passive VP Nomin of raced= gap θ-subject of raced= gap	[4]
	TOTAL =	26

(aus Hawkins [4])

References

[1] Demberg, Vera & Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2). 193–210.

[2] Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68(1). 1–76.

[3] Gibson, Edward. 2000. The dependency locality theory: a distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (eds.), *Image, language, brain: papers from the first mind articulation project symposium*, 95–126. Cambridge, MA: MIT Press.

[4] Hawkins, John A. 2014. Major contributions from formal linguistics to the complexity debate. In Frederick J. Newmeyer & Laurel B. Preston (eds.), *Measuring grammatical complexity*, 14–36. Oxford: Oxford University Press.