Mathematische Grundlagen III

Informationstheorie in der Linguistik und Psycholinguistik

Prof Dr. Matthew Crocker

Universität des Saarlandes

2. Juli 2015

Informationstheorie in der Psycholinguistik

Entropie erlaubt uns zu messen, wie viel Information übertragen wird.

Psycholinguistische Perspektive:

- Ist Verarbeitungsaufwand dieser Information proportional zur Menge der übertragenen Information?
- d.h.: können wir Verarbeitungsschwierigkeit die durch Sprache entsteht (z.B. einen schwer verständlichen Satz) mit informationstheoretischen Maßen messen?

Informationstheorie in der Linguistik

In der Entropievorlesung haben wir gesehen, wie wir einen Huffman Code finden können, sodass wir eine bestimmte Information möglichst effizient, d.h. mit einem kurzen Code übertragen können.

Linguistische Perspektive:

- Wie sieht es mit natürlicher Sprache aus? Ist natürliche Sprache ein effizienter Code?
- Ambiguität: bug oder feature?
  Was ist die kommunikative Funktion von Ambiguität?

Informationstheorie in der Linguistik

In der Entropievorlesung haben wir gesehen, wie wir einen Huffman Code finden können, sodass wir eine bestimmte Information möglichst effizient, d.h. mit einem kurzen Code übertragen können.

Linguistische Perspektive:

- Wie sieht es mit natürlicher Sprache aus? Ist natürliche Sprache ein effizienter Code?
  - **Psycholinguistische Zwischenfrage:**
    Sind Menschen in der Lage, diesen Code effektiv zu nutzen um "optimal" kommunizieren?
    (optimal = ausgewogen zwischen Effizienz und ausreichend Redundanz um verstanden zu werden.)
- Ambiguität: bug oder feature?
  Was ist die kommunikative Funktion von Ambiguität?

Inhaltsverzeichnis

Inhaltsverzeichnis

Informationstheorie in der Psycholinguistik

Informationstheorie erlaubt uns zu messen, wie viel Information ein Wort gegeben den Kontext enthält.
→ Bedingte Entropie an einem bestimmten Punkt (für konkrete x und y).

Bedingte Entropie

Wieviel Information wird benötigt um X mitzuteilen, wenn Y schon bekannt ist?

$$H(Y|X) = -\sum_x \sum_y p(x,y) \log_2 p(y|x)$$

Bedingte Entropie für ein bestimmtes Wort y in Kontext $x_1 ... x_n$:

$$H(y|x_1..x_n) = -\log_2 p(y|x_1..x_n)$$

- In psycholinguistischer Literatur wird diese bedingte Entropie eines bestimmten Wortes gegeben den Kontext als **Surprisal** bezeichnet.
- **Hypothese**: Verarbeitungsschwierigkeiten eines Menschen beim Verstehen von Sprache ist proportional zu Surprisal.

Interpretation von Surprisal

Was sollte Surprisal mit Verarbeitungsschwierigkeit zu tun haben?

$$Surprisal(y) = -\log_2 p(y|x_1..x_n)$$

- **Integrative Sichtweise:** zu jedem Zeitpunkt gibt es eine Reihe von möglichen Interpretationen eines Satzes. Surprisal misst die Anstrengung die nötig ist, um Interpretationen, die gegeben ein neues Wort nicht mehr möglich sind, zu eliminieren.
- **Prädiktive Sichtweise:** Menschen haben Erwartungen an die Sätze die sie hören oder lesen, d.h. sie erwarten gewisse Strukturen oder Wörter die gut zu dem bisher verarbeiteten passen. Wenn diese Erwartungen nicht erfüllt werden, entsteht Verarbeitungsschwierigkeit.

Surprisal

Surprisal

$$\text{Surprisal}_{k+1} = -\log P(w_{k+1}|w_1 \cdots w_k)$$

Surprisal kann auf verschiedene Art und Weise abgeschätzt werden:

- n-gram Surprisal (Beispiel für $n = 4$):
  $\text{Surprisal}(w_i) = -\log_2 p(w_i|w_{i-3}, w_{i-2}, w_{i-1})$
- syntactic Surprisal:

$$\begin{aligned} S_{k+1} &= -\log P(w_{k+1}|w_1 \cdots w_k) \\ &= -\log \frac{P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} \\ &= \log P(w_1 \cdots w_k) - \log P(w_1 \cdots w_{k+1}) \\ &= \log \textstyle\sum_T P(T, w_1 \cdots w_k) - \log \sum_T P(T, w_1 \cdots w_{k+1}) \end{aligned}$$

Syntactic Surprisal:

Syntactic Surprisal

$$S_{k+1} = -\log P(w_{k+1}|w_1 \cdots w_k)$$

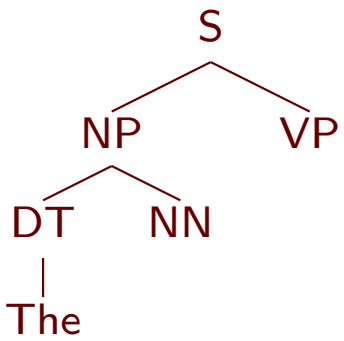
$$= \underbrace{\log \sum_T P(T, w_1 \cdots w_k)}_{\text{prefix probability}(w_k)} - \underbrace{\log \sum_T P(T, w_1 \cdots w_{k+1})}_{\text{prefix probability}(w_{k+1})}$$

Prefix Probability bei Wort w_k:

Summe über die Wahrscheinlichkeit aller Bäume T, die Worte $w_1..w_k$ überspannen.

$\rightarrow$ Diese Wahrscheinlichkeit können wir z.B. mit einer PCFG abschätzen.

Beispiel syntaktischer Surprisal



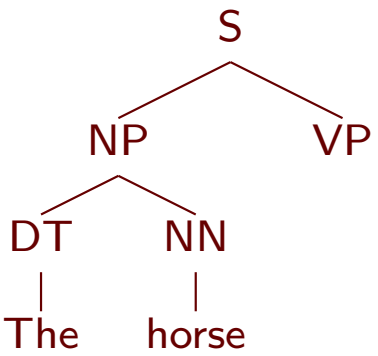
Surprisal berechnen:

- Prefix probabilities:
  $prefprob_{w_n} = -log \sum_T p(T|w_1 \ldots w_n)$
- Surprisal s of word w_n:
  $s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S $\rightarrow$ NP VP	$p = 1.0$
NP $\rightarrow$ DT NN	$p = 0.87$
NP $\rightarrow$ NP VP	$p = 0.13$
VP $\rightarrow$ VBD	$p = 0.07$
VBD $\rightarrow$ raced	$p = 0.0005$
VBN $\rightarrow$ raced	$p = 0.000001$
DT $\rightarrow$ the	$p = 0.7$
$\vdots$	$\vdots$

Beispiel syntaktischer Surprisal



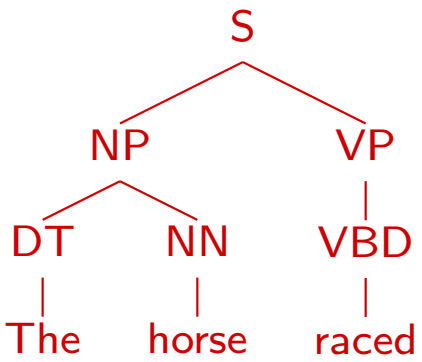
Surprisal berechnen:

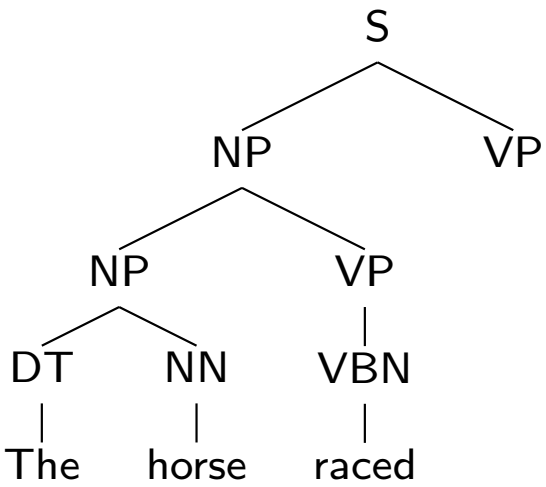
- Prefix probabilities:
  $prefprob_{w_n} = -log \sum_T p(T|w_1 \ldots w_n)$
- Surprisal s of word w_n:
  $s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S → NP VP	$p = 1.0$
NP → DT NN	$p = 0.87$
NP → NP VP	$p = 0.13$
VP → VBD	$p = 0.07$
VBD → raced	$p = 0.0005$
VBN → raced	$p = 0.000001$
DT → the	$p = 0.7$
⋮	⋮

Beispiel syntaktischer Surprisal





raced kommt wesentlich häufiger als simple past (VBD) vor, als als past participle (VBN) Form.

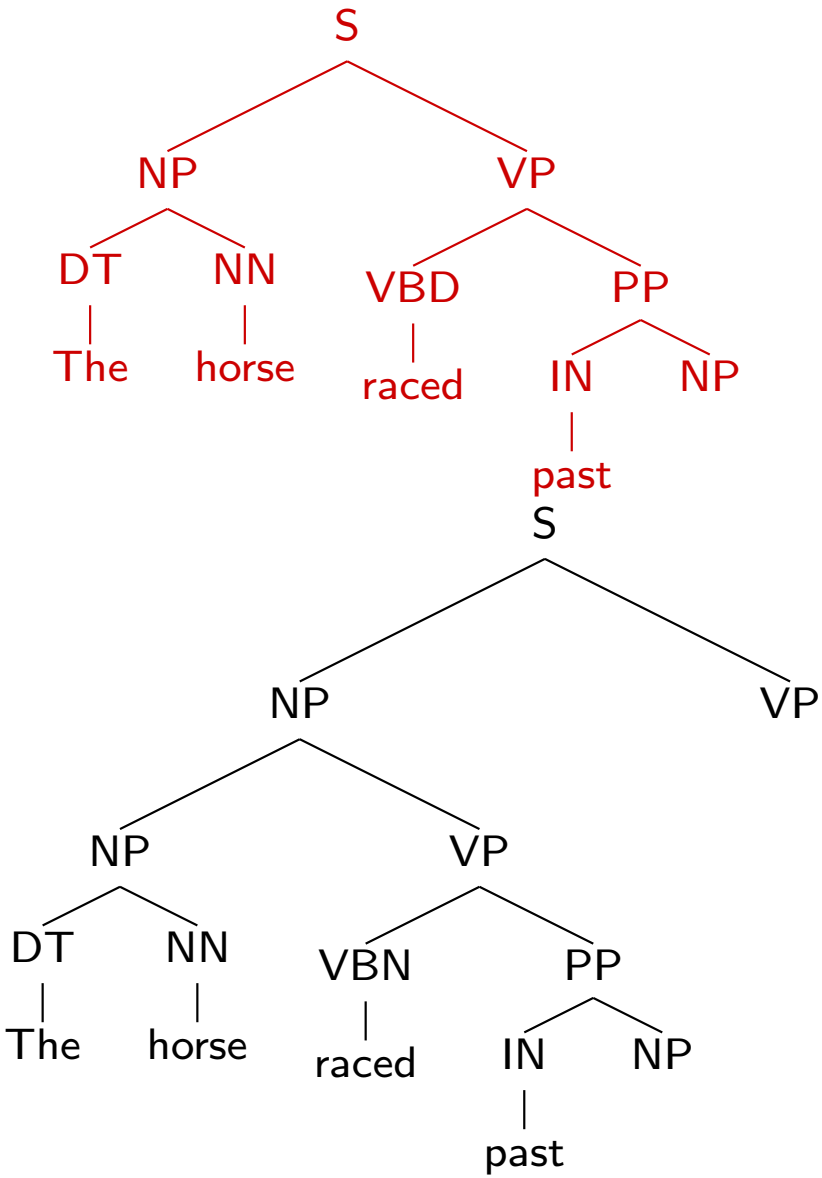
Surprisal berechnen:

- Prefix probabilities:
  $$prefprob_{w_n} = -log \sum_T p(T|w_1 \dots w_n)$$
- Surprisal s of word w_n:
  $$s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S $\rightarrow$ NP VP	$p = 1.0$
NP $\rightarrow$ DT NN	$p = 0.87$
NP $\rightarrow$ NP VP	$p = 0.13$
VP $\rightarrow$ VBD	$p = 0.07$
VBD $\rightarrow$ raced	$p = 0.0005$
VBN $\rightarrow$ raced	$p = 0.000001$
DT $\rightarrow$ the	$p = 0.7$
$\vdots$	$\vdots$

Beispiel syntaktischer Surprisal



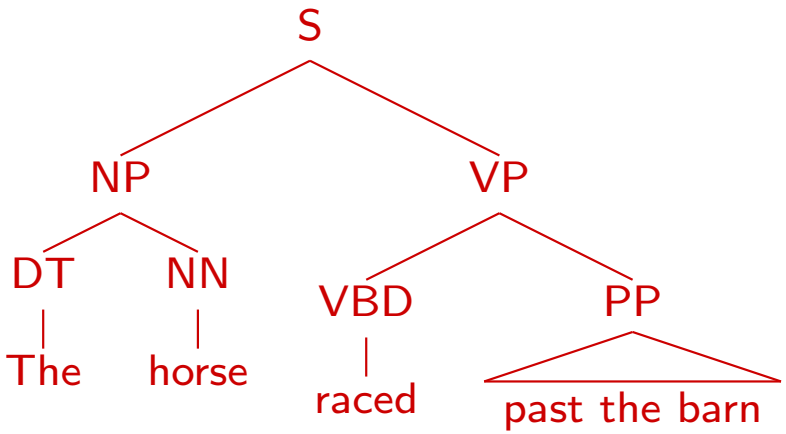
Surprisal berechnen:

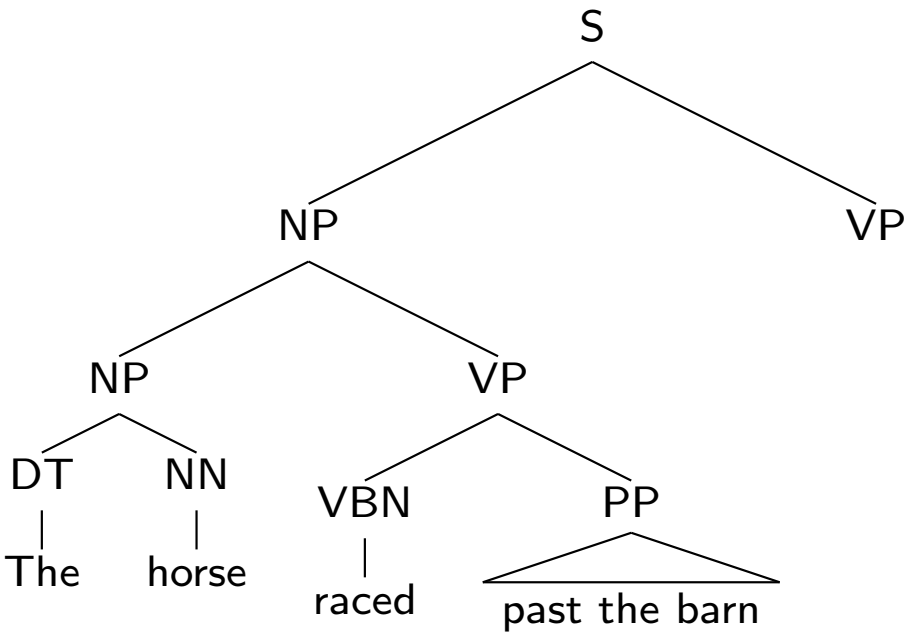
- Prefix probabilities:
  $$prefprob_{w_n} = -log \sum_T p(T|w_1 \ldots w_n)$$
- Surprisal s of word w_n:
  $$s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S → NP VP	$p = 1.0$
NP → DT NN	$p = 0.87$
NP → NP VP	$p = 0.13$
VP → VBD	$p = 0.07$
VBD → raced	$p = 0.0005$
VBN → raced	$p = 0.000001$
DT → the	$p = 0.7$
⋮	⋮

Beispiel syntaktischer Surprisal





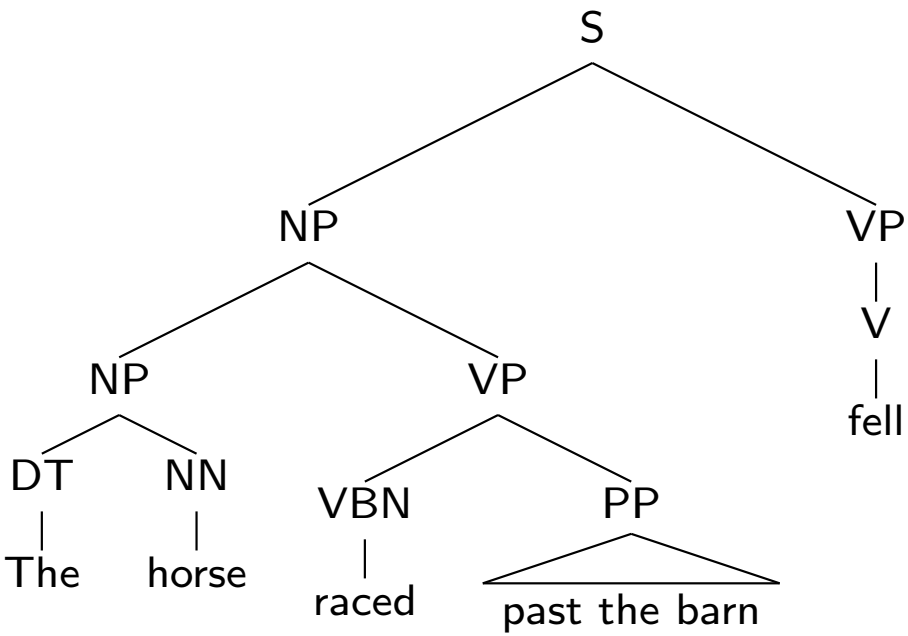
Surprisal berechnen:

- Prefix probabilities:
  $$prefprob_{w_n} = -log \sum_T p(T|w_1 \ldots w_n)$$
- Surprisal s of word w_n:
  $$s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S → NP VP	$p = 1.0$
NP → DT NN	$p = 0.87$
NP → NP VP	$p = 0.13$
VP → VBD	$p = 0.07$
VBD → raced	$p = 0.0005$
VBN → raced	$p = 0.000001$
DT → the	$p = 0.7$
⋮	⋮

Beispiel syntaktischer Surprisal



Surprisal berechnen:

- Prefix probabilities:

$$prefprob_{w_n} = -log \sum_T p(T|w_1 \ldots w_n)$$

- Surprisal s of word w_n:

$$s_{w_n} = prefprob_{w_n} - prefprob_{w_{n-1}}$$

Beispiel PCFG:

Regel	**Wahrscheinlichkeit**
S → NP VP	$p = 1.0$
NP → DT NN	$p = 0.87$
NP → NP VP	$p = 0.13$
VP → VBD	$p = 0.07$
VBD → raced	$p = 0.0005$
VBN → raced	$p = 0.000001$
DT → the	$p = 0.7$
⋮	⋮

Surprisal – Modelling Results (1)

Surprisal des Gartenpfad-Satzes

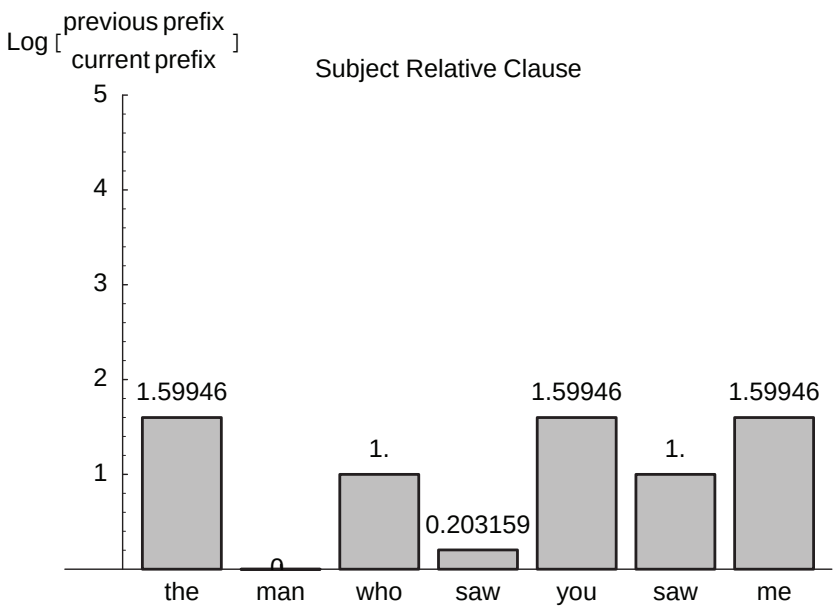
[Hale, 2001]

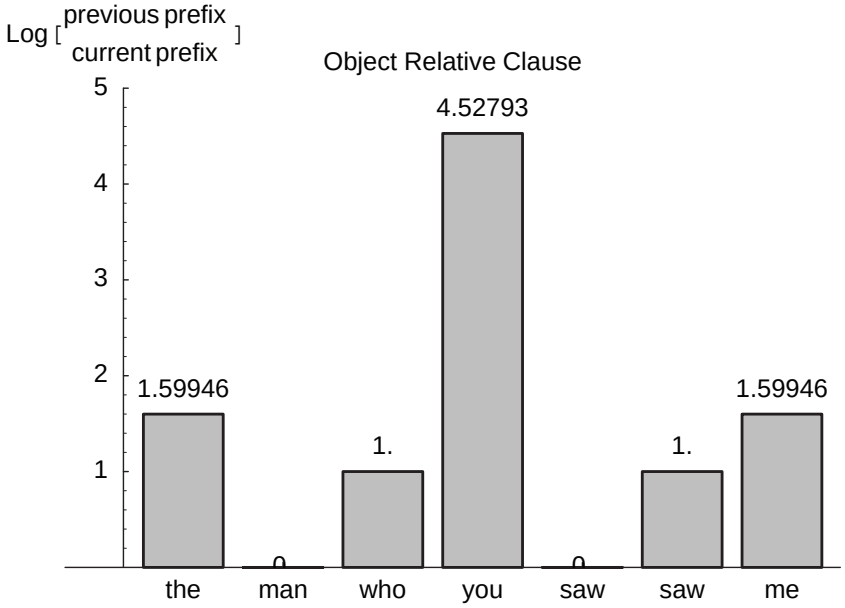
Surprisal – Subjekt- und Objektrelativsätze

Hintergrund

- Subjektrelativsätze sind leichter zu verstehen als Objektrelativsätze

Surprisal Vorhersagen





Inhaltsverzeichnis

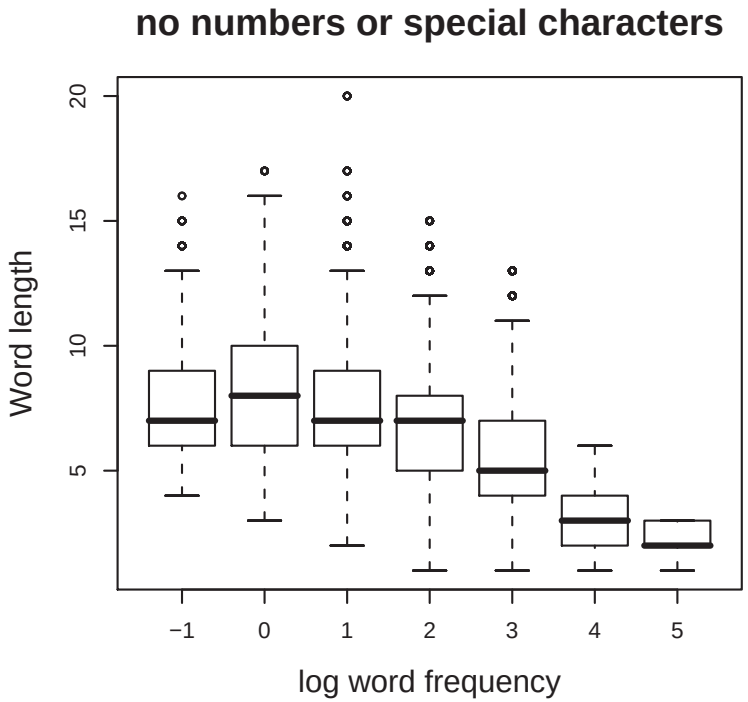
Ist Sprache ein effizienter Code?

Erinnern wir uns an Codedesign mit Huffman Code Trees.

- Ziel ist, einen effizienten Code zu entwickeln, damit wir unsere Information in möglichst kurze Nachrichten verpacken können.
- Dabei ist die Idee, häufige Ereignisse mit kürzeren Codes zu kodieren, damit die Länge einer durchschnittlichen Nachricht minimiert wird.
- **Ist natürliche Sprache so ein effizienter Code?**

Frequenz und Länge

Wir wissen schon, dass Frequenz und Länge miteinander korrelieren.



G. Zipf hat dies schon in den 1930er Jahren beobachtet, und dies im "Principle of Least Effort" formuliert. (Das war also sogar vor Shannon.)

Aber erinnern wir uns nochmal an die Entropie-Vorlesung:

- Je besser unser Sprachmodell, desto geringer die Entropie, und desto bessere Codes sollten wir konstruieren können.

  Beispiel: vereinfachtes Polynesisch bei dem wir nur Buchstabenfrequenzen kennen vs. vereinfachtes Polynesisch bei dem wir Buchstabenfrequenzen im Silbenkontext kennen.

- Die Beobachtung, dass Länge mit Frequenz korreliert, nimmt aber ein sehr simples Sprachmodell an (keine Kontextinformation).
- Müsste in einem System mit gutem Codedesign nicht die Vorhersagbarkeit eines Wortes eine viel stärkere Korrelation mit der Codelänge haben, als Wortfrequenz???

Aber erinnern wir uns nochmal an die Entropie-Vorlesung:

- Je besser unser Sprachmodell, desto geringer die Entropie, und desto bessere Codes sollten wir konstruieren können.

  Beispiel: vereinfachtes Polynesisch bei dem wir nur Buchstabenfrequenzen kennen vs. vereinfachtes Polynesisch bei dem wir Buchstabenfrequenzen im Silbenkontext kennen.

- Die Beobachtung, dass Länge mit Frequenz korreliert, nimmt aber ein sehr simples Sprachmodell an (keine Kontextinformation).
- Müsste in einem System mit gutem Codedesign nicht die Vorhersagbarkeit eines Wortes eine viel stärkere Korrelation mit der Codelänge haben, als Wortfrequenz???

Fragestellung

Hat die Vorhersagbarkeit eines Worte eine starke Korrelation mit der Wortlänge? Das können wir mithilfe eines Korpus testen.

- Wir ermitteln die durchschnittliche Vorhersagbarkeit eines Wortes w gegeben seinen Kontext c

$$\sum_{c} P(W = w|C = c) * -\log P(W = w|C = c)$$

Fragestellung

Hat die Vorhersagbarkeit eines Worte eine starke Korrelation mit der Wortlänge? Das können wir mithilfe eines Korpus testen.

- Wir ermitteln die durchschnittliche Vorhersagbarkeit eines Wortes w gegeben seinen Kontext c

$$\underbrace{\sum_{c} P(W = w|C = c)}_{\text{Gewichtung}} * \underbrace{-\log P(W = w|C = c)}_{\text{Surprisal}}$$

In der Praxis

Korpuswahl:

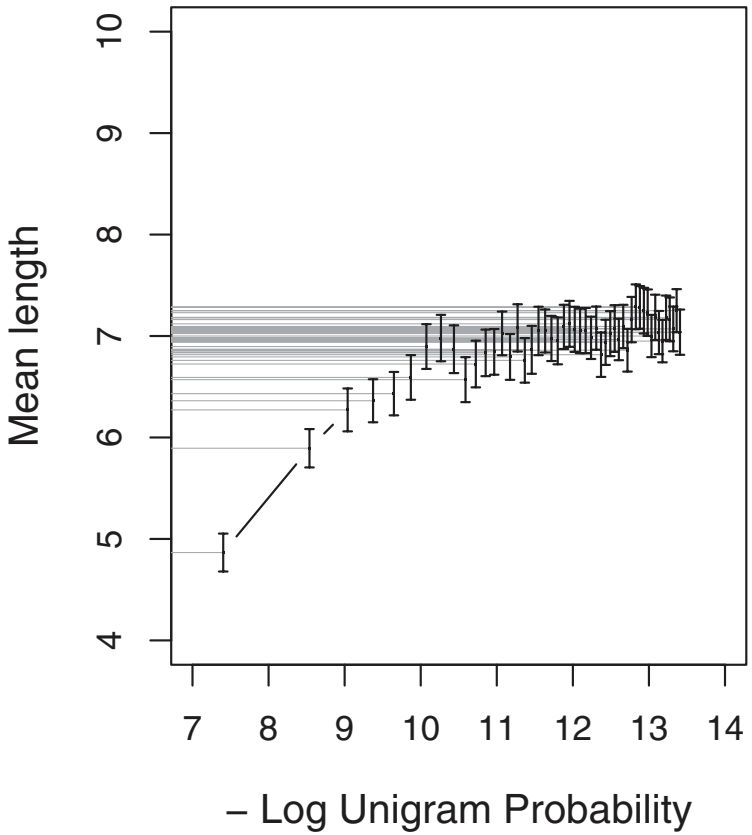
- Google n-grams (500 Millarden Wörter)

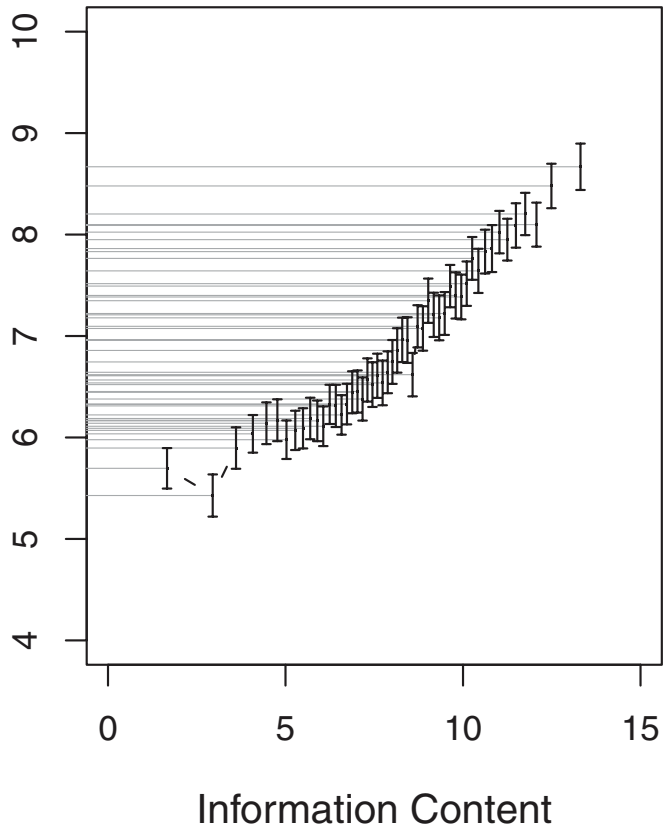
Vorhersagbarkeit eines Wortes:

- n-gram Surprisal
- kein Smoothing nötig

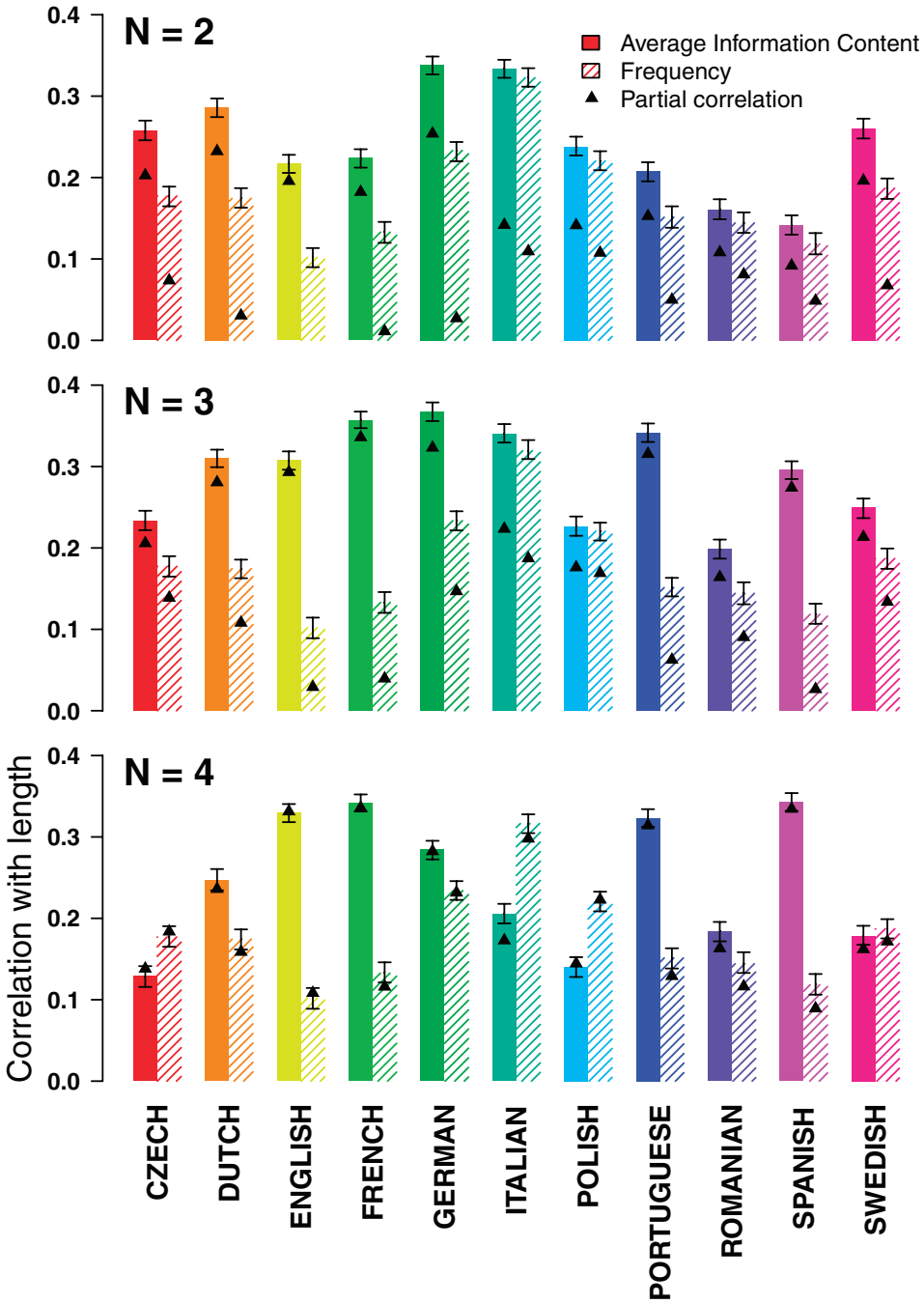
Vorhersagbarkeit vs. Frequenz

Korrelation von Frequenz und Länge vs. Informationsgehalt (n-gram Surprisal) und Länge





(hohe Frequenz = kleine -Log Unigram Probability)



- Informationsgehalt kann Wortlänge besser vorhersagen als Frequenz.
- Dies gilt für alle untersuchten Sprachen
- **Sprache ist ein guter Code:** Wortlängen sind optimiert für effiziente Kommunikation.

Warum ist gutes Codedesign für menschliche Sprache sinnvoll?

- Mit einem guten Code wird Information mit annähernd konstanter Rate übertragen.
- Leicht vorhersagbare Information wird kurz kodiert, größere Mengen von Information werden länger kodiert.
- So kann die Kanalkapazität von gesprochener Sprache oder dem kognitiven System gut genutzt werden.

Uniform Information Density

- Sprache ist so gewachsen, dass die Information mit recht gleichmäßiger Rate übertragen wird.
- Können Menschen sie dynamisch so benutzen, dass diese Rate weiter optimiert wird?
  $\longrightarrow$ Uniform Information Density Hypothesis

Uniform Information Density

Können Menschen sie dynamisch so benutzen, dass diese Rate weiter optimiert wird?

- n-gram Surprisal hat einen Effekt auf Wortwahl und Strukturwahl:
  - Levy and Jaeger (2007): reduction of optional that-complementizers in English inversely proportional to 3-gram surprisal.
  - Frank and Jaeger (2008): UID account can explain reduction of be/have/not in English.
- n-gram surprisal hat einen Effekt auf Silbenlänge:
  - Aylett and Turk (2006): n-gram probability of a syllable has inverse relationship to syllable duration in speech.
- Worte werden schneller ausgesprochen, wenn sie gegeben das vorige/folgende Wort eine hohe Wahrscheinlichkeit haben.
  - Jurafsky et al. (2001): corpus study, but doesn't use syntactic surprisal.
- Wörter an denen hoher syntaktischer Surprisal auftritt werden langsamer ausgesprochen (Demberg et al., 2012)

Uniform Information Density - Demberg et al., 2012

Korpus

- AMI meeting corpus:
  - has fine-grained word durations.
  - "correctly" spelled transcriptions: necessary for parsing.
  - speaker turns, disfluencies annotated
- Focus group meetings for designing remote control.
- Over 1M words. Native and non-native English speakers.

Korpus

Natürlich sind “echte” Korpusdaten immer sehr verrauscht. Um damit arbeiten zu können, muss man erst ein bisschen aufräumen.

Vorverarbeitungsschritte AMI corpus:

- Remove non-words such as “hmm” and “uhm”.
- Remove incomplete words or incorrectly transcribed words (“recogn”). (loss rate 2.9% of word types)
- Remove incomplete sentences.

Uniform Information Density

Wir dürfen aber nicht Wortdauern direkt mit syntaktischem Surprisal korrelieren.

- Andere Faktoren sind mit syntaktischem Surprisal korreliert:
- Laenge des Wortes
- Frequenz
- n-gram Surprisal

Wir müssen erst alle diese anderen Faktoren rausrechnen, um dann zu sehen, ob syntaktischer Surprisal danach noch Vorhersagekraft hat.

Resultate:
Syntaktischer Surprisal kann Wortdauern besser vorhersagen als Frequenz oder n-gram Surprisal.

Uniform Information Density

Wir dürfen aber nicht Wortdauern direkt mit syntaktischem Surprisal korrelieren.

- Andere Faktoren sind mit syntaktischem Surprisal korreliert:
- Laenge des Wortes
- Frequenz
- n-gram Surprisal

Wir müssen erst alle diese anderen Faktoren rausrechnen, um dann zu sehen, ob syntaktischer Surprisal danach noch Vorhersagekraft hat.

Resultate:
Syntaktischer Surprisal kann Wortdauern besser vorhersagen als Frequenz oder n-gram Surprisal.

Inhaltsverzeichnis

Ambiguität (Piantadosi et al., 2012)

- Taucht auf allen Ebenen der linguistischen Analyse auf.
  - Wörter haben mehrere Bedeutungen
  - Morpheme können ambig sein (-s in Peter**s**, CD**s**)
  - unterschiedliche Worte können gleich ausgesprochen werden
    arm, Arm; konnten, Konten; küsste, Küste; laichen, Leichen; Trend, trennt; misst, Mist; Waagen, wagen, Wagen...
  - syntaktische Ambiguität (z.B. Attachment)
  - semantische Ambiguität (z.B. Skopus)
- Wäre es nicht viel besser für Kommunikation, wenn Sprache **eindeutig** wäre???

Warum Ambiguität?

Hypothese:
Ambiguität erlaubt es uns, einen kurzen und einfachen Code zu benutzen.

- Kontext (Situation, Weltwissen, restliche Äußerung) disambiguiert Sprache sehr häufig.
- Ein unambiges Sprachsignal wäre dann teilweise redundant.
- Ambiguität erlaubt es uns, Sprachsignale, die einfach zu produzieren / verstehen sind, mehrfach zu benutzen.

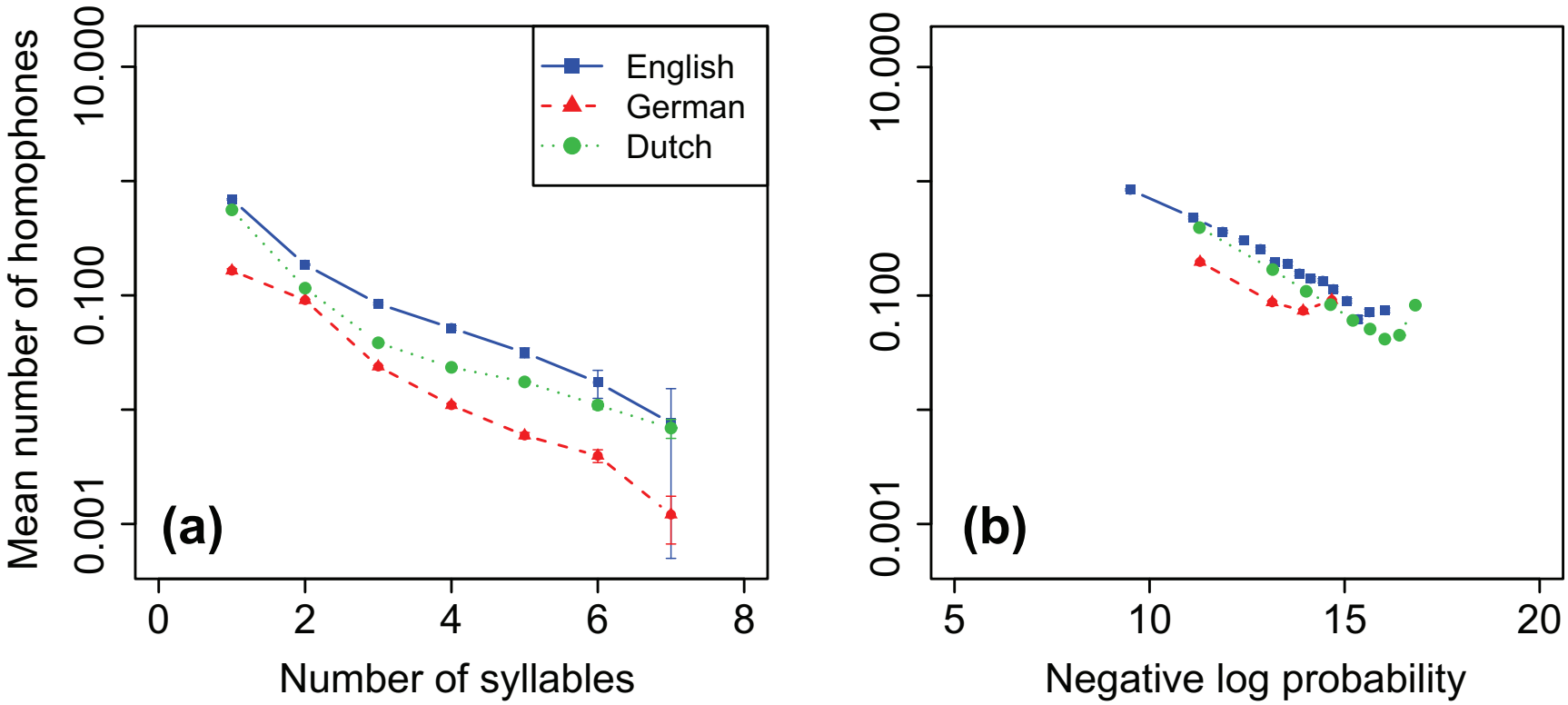
Experiment

Fragestellung: Wie ist Ambiguität verteilt?

- Falls Ambiguität einfach ein zufälliger “bug” ist, würden wir nicht erwarten, dass kurze oder häufige Wörter/Silben/Phoneme stärker mehrdeutig sind als lange oder seltene Wörter/Silben/Phoneme.
- Falls Ambiguität aber ein Mittel ist, um den Code zu optimieren (im Sinne von leicht zu produzieren und zu verstehen), dann sollten wir beobachten können, dass kurze Wörter / Silben / Phoneme stärker mehrdeutig sind, um diese leichten kurzen Einheiten wiederverwenden zu können.

Piantadosi et al., 2012

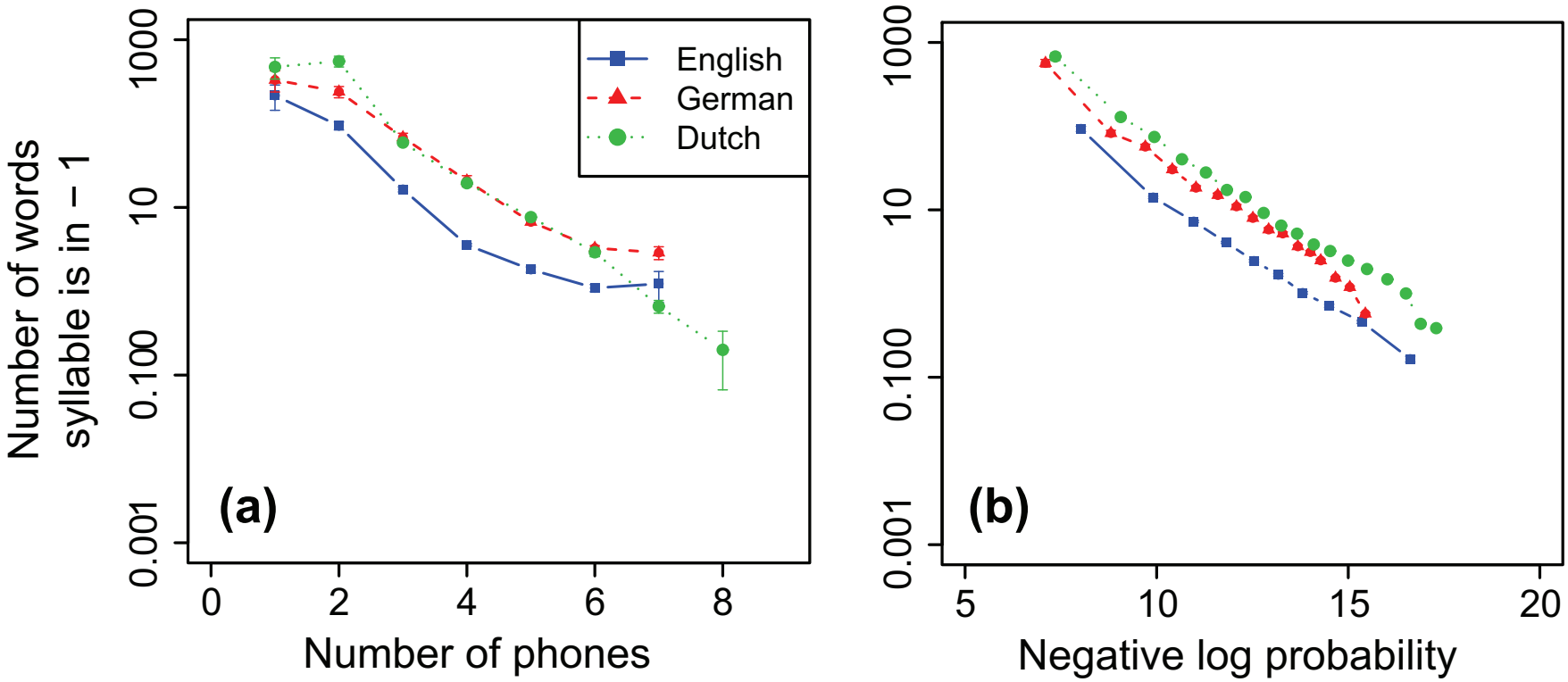
Analyses von Homophonen (Wörter die gleich klingen).



(geringe negative log probability = hohe Frequenz)

Piantadosi et al., 2012

Analyse von Silben.



(geringe negative log probability = hohe Frequenz)

Zusammenfassung

Ambiguität als Feature!

- Hilft uns in kürzeren Codes zu kommunizieren.
- Hilft uns, einfache und kurze Codes wiederzuverwenden.
- Echte Ambiguität (d.h. Mehrdeutigkeit sogar wenn voller Kontext präsent ist) ist selten

Zusammenfassung

Surprisal als Maß für menschliche Verarbeitungsschwierigkeit.

- Surprisal = punktuelle relative Entropie
- Surprisal quantifiziert wie unerwartet ein Wort gegeben einen Kontext ist.

Sprache als effizienter Code?

- hochfrequente Wörter sind kürzer als selten Wörter.
- Darüber hinaus: Wörter die oft leicht vorhersagbar sind, sind kurz.
- D.h.: menschliche Sprache ist informationstheoretisch effizient, information wird mit relativ konstanter Rate übertragen.
- Menschen optimieren diese Rate weiterhin durch Wortwahl, Strukturwahl, Aussprachelänge

Ambiguität – bug oder feature?

- Ambiguität ist ein Feature.
- Ambiguität hilft, leicht zu produzierende und leicht zu verstehende kurze Worte / Silben / Laute wiederzuverwenden.
- Dabei ist echte Ambiguität mit Kontext eher selten; echte