



# Korpuslinguistik & das Web

Proseminar „Web-basierte Techniken in der  
Computerlinguistik“  
WS 2012 / 2013  
Michaela Regner



## „Sessel-Linguisten“ vs. Korpuslinguisten



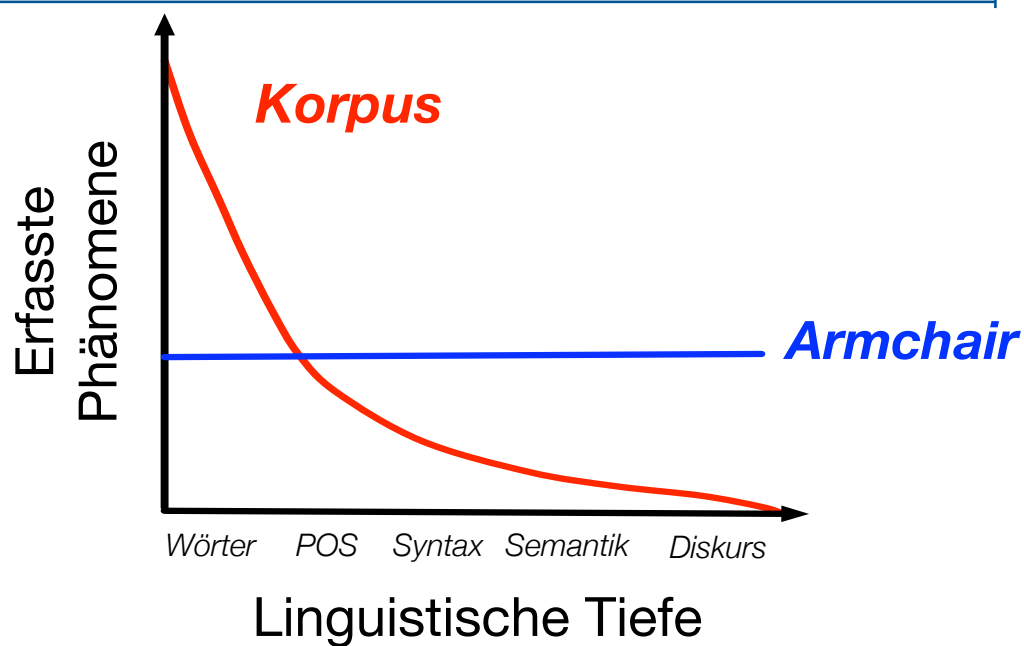
Kompetenz



Performanz



## Korpus vs. Armchair



## Übersicht

- Korpuslinguistik - Grundlagen
  - Korpora
  - Annotation
- Korpusanalyse
- Das Web als Korpus
- Crowdsourcing: Spezialfall Wikipedia
- Crowdsourcing für mehr Annotation



# Korpora

---

- im Prinzip: jede Sammlung von Text
- für linguistische Anwendungen im Idealfall
  - repräsentativ (nicht nur Fragesätze aus Staubsauger-FAQs z.B.)
  - maschinenlesbar
  - eine Standard-Referenz für unterschiedliche Benutzer



## Korpora - Charakteristika (1)

---

- Größe
- Sprachen und Alignierung: Textsprache(n), einsprachig vs. mehrsprachig, vergleichbar vs parallel
- Sprachmodus: geschrieben vs. gesprochen
- Textarten (z.B. Zeitungen, Telefongespräche, Briefe,...)
- Textdomänen (z.B. Wirtschaft, Liebesbriefe, ...)



## Korpora - Charakteristika (2)

---

- Balanciertheit: homogen vs. inhomogen  
balanciert vs. unbalanciert
- Entstehungsdatum der Texte
- Annotation:
  - Reintext vs. annotierter Text
  - Art der Annotation



## Annotation - Prinzipien

---

- zusätzliche (hier: linguistische) Information im Korpus
- Annotationsmaximen (Leech 1993):
  - entfernen- und extrahierbare Annotation
  - Nutzer sollte Annotationsrichtlinien bekommen
  - Einkalkulieren von Fehlern (aber auch möglichem Nutzen)
  - größtmöglicher, theorieneutraler Konsens



# Annotation - Format

- Oft XML-Varianten (vgl. HTML):

*The dog barks.*

„stand-off“:

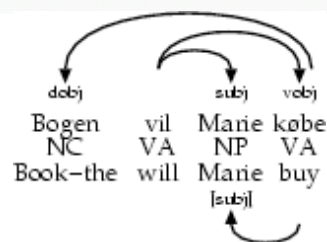
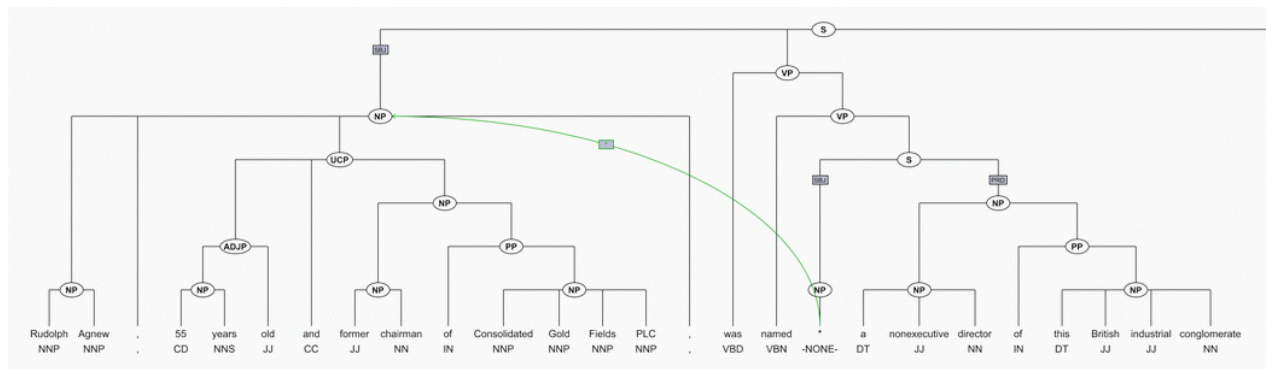
```
<sentence>
  <phrase type="NP">
    <word ind="1" pos="det"/>
    <word ind="2" pos="N"/>
  </phrase>
  <phrase type="VP">
    <word ind="3" pos="VI"/>
  </phrase>
</sentence>
```

„inline“:

```
<sentence>
  <phrase type="NP">
    <word pos="det">the</word>
    <word pos="N">dog</word>
  </phrase>
  <phrase type="VP">
    <word pos="VI">barks</word>
  </phrase>
</sentence>
```

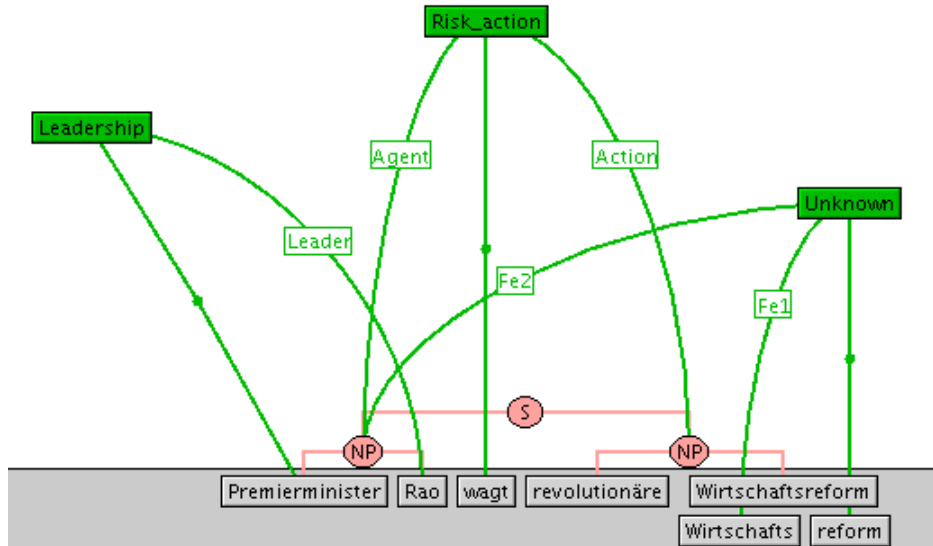


# Annotation - Beispiele: Baumbanken (Syntax)

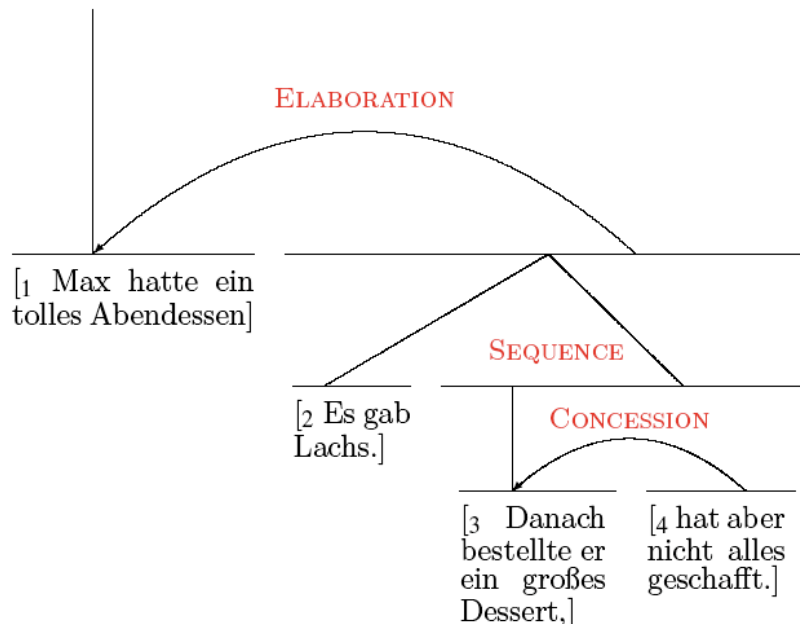




# Annotation - Beispiele: Semantische Rollen (SALSA)



# Annotation - Beispiele: Diskurs-Relationen





# Annotation - woher?

---

- manuell annotierte Referenz-Korpora
  - Annotations-Tools, benutzt von
    - unzähligen Hiwis / Studenten / Doktoranden
    - teuer, zeit- und arbeitsintensiv
  - automatische Tools, die Korpora annotieren - die meisten lernen aus Referenzkorpora
  - schwierigere Annotations-Aufgabe → kleinere Korpora → seltenere / unzuverlässigere Tools



# Annotation - Qualitätsbeurteilung

---

- für manuell annotierte Korpora:
  - Übereinstimmung der Annotatoren
  - Übereinstimmung der Annotatoren mit „Gold Standard“
- für automatische Tools:
  - Qualität des Tools auf dem Korpus (Übereinstimmung mit Gold Standard)
  - oft unterschiedliche Performanz für unterschiedliche Domänen!



# Korpusanalyse - Grundlagen (1)

- abstrakt: (zielgerichtetes) Zählen und Vergleichen von Dingen
- „Dinge“ sind hier Worte, Lemmata, Satzbaume, ...
- einfache Beispiele fürs Zählen:
  - Wörter (Worthäufigkeiten, Tokens pro Type)
  - N-Gramme: Häufigkeiten von Wortkombinationen der Länge N



# Korpusanalyse - Grundlagen (2)

- Einfache Beispiele fürs Vergleichen / Sammeln:
  - Konkordanzen: ein Wort im Kontext

Burschenschaft »Arminia« auch die Karlsruher Händel-Festspiele. Als Girokonto für alle Azubis, Schüler, und richtet sich an Seminar richtet sich vorzugsweise an 2000 (Programme for International	Studenten Student Studenten Studenten Studenten Student	der Technischen Hochschule in in Halle und Göttingen von den usw. für 0 DM Gebühren. Außerdem in den ersten Semestern. des Grundstudiums im 1. oder 3. Assessment) ergeben, dass die
--	--	--

- Pattern-Suche: Was machen Kinder meistens?

<i>Kinder antworten meistens</i>	auf schwedisch - aus
<i>Kinder verschwinden meistens</i>	in die Häuser hinein, es
<i>Kinder reagieren meistens</i>	noch instinktiv auf
<i>Kinder gehen meistens</i>	nur einmal die Woche in
<i>Kinder wachsen meistens</i>	in einem armen soziale
<i>Kinder fallen meistens</i>	durch ein verändertes





## Korpusanalyse - Anwendungen (1)

- (N-Gramm-)Sprachmodelle (für automatische Spracherkennung, maschinelle Übersetzung):

der Hund  laut

vergleiche:

p(„Hund dröhnt“)	p(„Hund schreit“)
p(„Hund bellt“)	
p(„wedelt laut“)	p(„sabbert laut“)
p(„bellt laut“)	



## Korpusanalyse - Anwendungen (2)

- allgemein: maschinelles Lernen zum automatischen Annotieren (besten Satz-Baum, beste POS-Tag-Sequenz...)
- Sammeln von (linguistischem) Wissen, z.B.
  - typische Objekte von bestimmten Verben
  - Häufigkeit von unterschiedlichen Bedeutungen eines Wortes



# Einfache Korpus-Analyse

- ist eine „Spaghetti Napoli“ eine Kollokation?
- Prinzip: überprüfe, ob Spaghetti und Napoli öfter zusammen auftreten, als per Zufall zu erwarten wäre
- Ein numerisches Maß, das Korrelationen misst, ist z.B. „Pointwise Mutual Information“ (PMI)
- auch eingesetzt für Pattern-Suche etc.

$$PMI = \log \frac{p(X, Y)}{p(X) \times p(Y)}$$

Wenn X und Y unabhängig sind gilt  
 $p(X, Y) = p(X) * p(y)$

- $PMI = 0$ : X ist **unabhängig** von Y
- $PMI > 0$ : X tritt **häufiger** mit Y auf als man bei Unabhängigkeit erwartet



# Einfache Korpus-Analyse - Beispiel

DeWac-Korpus: 1.4 x 10<sup>9</sup> Wörter  
 Spaghetti: 2448 Vorkommen  
 Napoli: 310 Vorkommen  
 Spaghetti Napoli: 12 Vorkommen

$$PMI = \log \frac{p(X, Y)}{p(X) \times p(Y)}$$

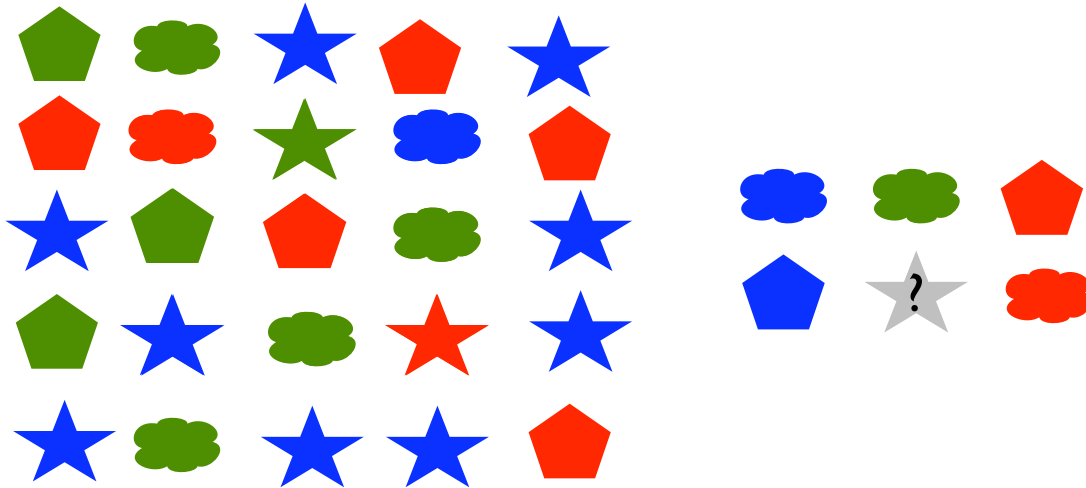
$p(\text{Spaghetti, Napoli}) \approx 12 / (1.4 \cdot 10^9)$   
 $p(\text{Napoli}) \approx 310 / (1.4 \cdot 10^9)$   
 $p(\text{Spaghetti}) \approx 2448 / (1.4 \cdot 10^9)$

Wahrscheinlichkeiten  
 abschätzen durch  
**relative Häufigkeiten**

$$PMI = \log \frac{p(\text{Spaghetti, Napoli})}{p(\text{Spaghetti}) \times p(\text{Napoli})} = 4.35$$



# Korpusanalyse - der „Flaschenhals“ (schematisch)



# Korpusanalyse - Datenmangel

- nie gesehene Wortsequenzen sind schwer oder gar nicht zu klassifizieren
- kleinere Korpora liefern weniger und unsichereres Wissen;  
Redundanz ist wichtig für gute Modelle
- manuell annotierte Korpora sind selten, teuer, meistens klein und aus speziellen Domänen



# Datenmangel?

---

roflcopter.

♪ Γ ( · o · ) ┘ ♪ L ( · o · ) Γ ♪



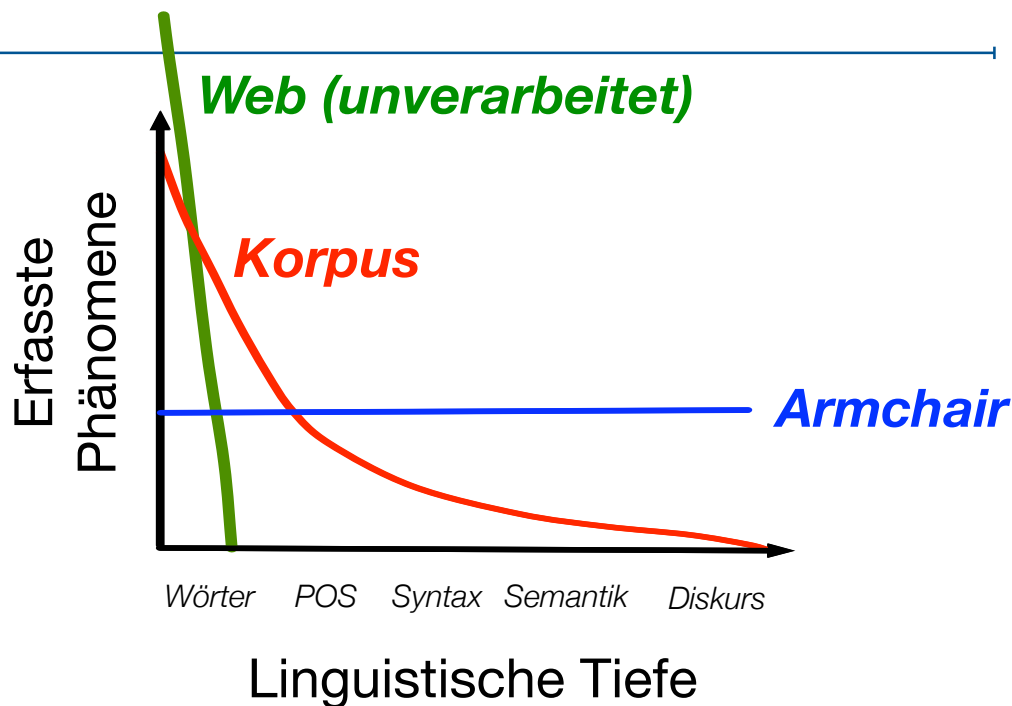
# Das Web als Korpus

---

- das Web ist eine Textsammlung, also ist es ein Korpus
- geschätzt  $45 \cdot 10^9$  indexierte Webseiten,  $10^{12}$  Wörter
- Charakteristika:
  - geschriebener und gesprochener Text
  - mehrsprachig, teilweise vergleichbar / parallel
  - (fast?) alle Textarten, -Domänen, ...



# Das Web als Korpus - Einordnung



# Das Web als Korpus - Suchmaschinen (1)

- Anzahl der Dokumenttreffer korrelieren direkt mit echten Wortfrequenzen (Keller 2003)
- einfaches Beispiel: wie schreibt man [ˌmɪsɪˈsɪpɪ] ?

„Mississippi“:	2 420 000	Treffer
„Misissippi“:	338 600	Treffer
„Mississippi“:	12 600 000	Treffer
„Mississippi“:	340 000 000	Treffer



# Das Web als Korpus - Suchmaschinen (2)

- leichter Zugriff auf „Snippets“
- einfaches Beispiel: Was bedeuten „T“ und „A“ in „E.T.A. Hoffmann“?

Google "Ernst \* \* Hoffmann" Suche [Erweiterte Suche](#)  
[Einstellungen](#)  
Suche:  Das Web  Seiten auf Deutsch  Seiten aus Deutschland

**Web**

[E.T.A. Hoffmann | Xlibris](#)  
Ernst Theodor Amadeus Hoffmann. Zitate von und über E.T.A. Hoffmann: Glimpf und Schimpf, Verstand und Überschwenglichkeit, Grauen and schallendes Gelächter ...  
[www.xlibris.de/Autoren/Hoffmann](http://www.xlibris.de/Autoren/Hoffmann) - 12k - [Im Cache](#) - [Ähnliche Seiten](#)

[E.T.A. Hoffmann - Biografie WHO'S WHO.](#)  
Ernst Theodor Amadeus Hoffmann, eigentlich Ernst Theodor Wilhelm Hoffmann, wurde am 24. ... Ernst Theodor Amadeus Hoffmann starb am 25. Juni 1822 in Berlin. ...  
[www.whoswho.de/templ/te\\_bio.php?PID=607&RID=1](http://www.whoswho.de/templ/te_bio.php?PID=607&RID=1) - 38k - [Im Cache](#) - [Ähnliche Seiten](#)

[Ernst Theodor Amadeus Hoffmann Biografie](#)  
Der deutsche Dichter Ernst Theodor Amadeus Hoffmann wird am 24.1.1776 in Königsberg geboren. Nach der Trennung der Eltern, der Vater ist Rechtsanwalt, ...  
[www.ernst-theodor-amadeus-hoffmann.de/](http://www.ernst-theodor-amadeus-hoffmann.de/) - 14k - [Im Cache](#) - [Ähnliche Seiten](#)



# Das Web als Korpus - Suchmaschinen (3)

- Zugriff auf Google etc. entweder
  - über APIs (DuckDuckGo!, Bing, Google bedingt) oder
  - über Links & Skripte:  
<http://www.google.com/search?q=%22suchbegriff1%20suchbegriff2%22>



# Das Web als Korpus - linguistische Tools / Ressourcen

---

- WebCorp: ein Konkordanz- als Suchmaschinen-Aufsatz  
<http://www.webcorp.org.uk/>
- Web as Corpus Toolkit (ausgelaufen):  
<http://www.drni.de/wac-tk/>
- Google 5-grams:  
<http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>



# Das Web als Korpus - ein paar Herausforderungen

---

- Webtexte sind nicht annotiert
- Webtexte darf jeder Verfassen, dadurch sind sie oft
  - ungrammatisch
  - linguistisch unbrauchbar (z.B. Preislisten)



# Das Web als Korpus - ein paar Herausforderungen

---

- Trefferzahlen beinhalten oft den selben Wortlaut mehrfach
- Trefferzahlen sind unpräzise Schätzungen
- Suchpräzision ist eingeschränkt (Interpunktion, Groß-/Kleinschreibung)



# Spezialfall Blogs & Twitter

---

- stark meinungsgeprägt
- großartige Quelle für Markt- / Trendforschung, oder allgemein “opinion mining”
- Rechtschreibung besonders brüchig
- besonders Twitter:
  - Abkürzungen, Emoticons, Links, \\_(ツ)\_/~, ...
- vollständige Sätze eher selten





# Crowdsourcing: Die Kraft der Masse

- Im Web gibt es zahllose Menschen, und jeder darf Schreiben, was er möchte
- Bsp. Wikipedia:
  - vielfältige Lexikonbeiträge
  - hohe Präzision durch das *Korrektiv der Masse*
  - *Annotation* - hier mit diversen Links



## Wikipedia als Korpus (1)



WIKIPEDIA  
The Free Encyclopedia

Navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

Search

Go Search

Information

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

Tools

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

Languages

- Cesky
- Deutsch
- Ελληνικά
- English

### Text corpus

From Wikipedia, the free encyclopedia

In **linguistics**, a **corpus** (plural *corpora*) or **text corpus** is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe.

A corpus may contain texts in a single language (*monolingual corpus*) or text data in multiple languages (*multilingual corpus*). Multilingual corpora that have been specially formatted for side-by-side comparison are called *aligned parallel corpora*.

In order to make the corpora more useful for doing linguistic research, they are often subjected to a process known as **annotation**. An example of annotating a corpus is **part-of-speech tagging**, or *POS-tagging*, in which information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of *tags*. Another example is indicating the **lemma** (base) form of each word. When the language of the corpus is not the user's, **interlinear glossing** is used to make the annotation bilingual.

Some corpora have further *structured* levels of analysis applied. In particular, some corpora are **parsed**. Such corpora are usually called **Treebanks** or **Parsed Corpora**. The analysis and annotation of such corpora is completely and consistently annotated means that these corpora are usually used for linguistic research. Other levels of linguistic structured analysis are possible, including annotation of syntactic structures.

Corpora are the main knowledge base in **corpus linguistics**. The analysis and processing of corpora are also the subject of much work in **computational linguistics**, **speech recognition** and **machine learning**. They are often used to train **hidden Markov models** for POS-tagging and other purposes. Corpora and their analysis are also used in **language teaching**.

Archaeological corpora

Text corpora are also used in the study of **historical documents**, for example in attempts to **decipher** ancient scripts, or in **Biblical scholarship**. Some archaeological corpora can be of such short duration that they provide a snapshot in time. One of the shortest corpora in time, may be the 15-30 year **Amarna letters** texts-(1350 BC). The *corpus* of an ancient city, (for

Links zu  
verwandten  
Begriffen

Links zu  
anderen  
Sprachen



## Wikipedia als Korpus (2)

---

- weitgehend grammatischer Fließtext
- (wikipedia-typische) Annotationen:
  - Links zur entsprechenden Seite in anderer Sprache
  - Kategorien (semantische Zuordnung?)
  - Links zu verwandten Begriffen (semantische Netze?)



## Wikipedia als Korpus (3)

---

- beschränkter Text-Typ: Lexikonartikel (aber viele Domänen, und Links zu externen Seiten mit ähnlicher Thematik)
- nur neutral gehaltene Sprache
- im Vergleich zum Web klein (aber im Vergleich zu anderen semantischen Ressourcen riesig)
- lexikon-typische sprachliche Besonderheiten (Artikel-Titel wird selten wiederholt, ... )



# Crowdsourcing

---

- nicht nur die Ergebnisse sind für Linguisten nutzbar, auch die Technik selbst
- Mechanical Turk: bezahlte “Microtasks” werden von riesigem Publikum erledigt
- Online-Spiele: Setzen auf freiwillige Beteiligung an “Games with a Purpose”
- In jedem Fall: Annotationen über das Web!
- Auch hier: Redundanz / Korrektiv der Masse



# Crowdsourcing: Mechanical Turk

---

- kommerzielle Plattform von Amazon
- jeder kann als “Requester” Tasks online stellen, die man übers Web machen kann
- jeder mit Amazon-Account kann als “Worker” solche Tasks lösen
- Organisation von Bezahlung etc. übernimmt Amazon, gegen 10% des Lohns
- beliebt für Annotationsaufgaben & Datensammlung



# Crowdsourcing: Mechanical Turk

[Rate the quality of computer-generated speech - Italian native only \(non-natives will be rejected\)](#) [View a HIT in this group](#)

**Requester:** [David Vazquez](#) **HIT Expiration Date:** Oct 29, 2012 (2 days 23 hours) **Reward:** \$0.35  
**Time Allotted:** 60 minutes **HITs Available:** 6

[Evaluate these statements \(10 minute survey\)](#) [View a HIT in this group](#)

**Requester:** [CrowdFlower](#) **HIT Expiration Date:** Nov 1, 2012 (6 days 4 hours) **Reward:** \$0.34  
**Time Allotted:** 60 minutes **HITs Available:** 115

[Write a 450-700 word article](#) [View a HIT in this group](#)

**Requester:** [Jayavinoth](#) **HIT Expiration Date:** Dec 14, 2012 (7 weeks) **Reward:** \$0.30  
**Time Allotted:** 60 minutes **HITs Available:** 2

[Test a golf magazine website 14](#) [View a HIT in this group](#)

**Requester:** [Charles McLaughlin](#) **HIT Expiration Date:** Nov 25, 2012 (4 weeks 2 days) **Reward:** \$0.30  
**Time Allotted:** 60 minutes **HITs Available:** 1

[Personality and Facebook - North American participants only](#) [View a HIT in this group](#)

**Requester:** [Lydia Bickel](#) **HIT Expiration Date:** Nov 7, 2012 (1 week 5 days) **Reward:** \$0.30  
**Time Allotted:** 60 minutes **HITs Available:** 1



# Crowdsourcing: Online-Spiele

- Tasks in Spiele verpackt (hier: linguistische Tasks)
- Idee: wer freiwillig und motiviert spielt, arbeitet besser, billiger und "fairer"
- Herausforderung 1: "langweilige" Tasks zu einem motivierenden Spiel zu verarbeiten
- Herausforderung 2: gute Erfolgskriterien, die Spieler motivieren & schlechte Daten filtern



# Crowdsourcing: Online-Spiele

<http://www.gwap.com>

score 880 Verboosity It's common sense time 0:08

Bonus

the secret word is... madman.

middle?

clues

it is a lunatic

it is a type of insane person

it has six letters

it looks like  + submit

about the same size as

it is related to

guesses

person?

number?

always?

pass



## Zusammenfassung

- Grundlegendes zur Korpuslinguistik
- Korpora als Modelle der linguistischen Wirklichkeit
- größere Korpora sind näher an der Wirklichkeit (aber haben meist weniger tiefe Informationen)
- das Web ist das größte Korpus
- Crowdsourcing als Quelle für Annotationen (Wikipedia als Resource, oder eigene Tasks)