# Das Web als Korpus für die linguistische Forschung

Martin Volk
Universität Zürich

## Beispiel für linguistische Ressource im Web

- Textverarbeitung WinEdt (ein Editor für LaTeX-Dokumente)
- mit Zugriff auf das online DE-EN LEO-Wörterbuch

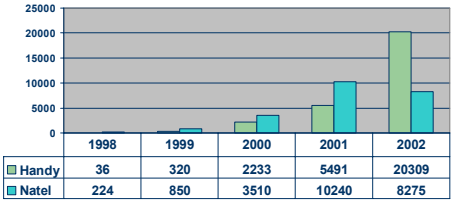


## Diachrone Studie: *Handy* vs. *Natel*

- Suche eingeschränkt
  - auf Deutsch
  - auf Web-Adressen in der Schweiz
    (d.h. auf die Domäne '.ch,)
- AltaVista-Häufigkeiten *(number of pages found)*
  - *Handy*: 28'417 (letztes Jahr: 23'792)
  - *Natel*: 23'156 (letztes Jahr: 15'165)



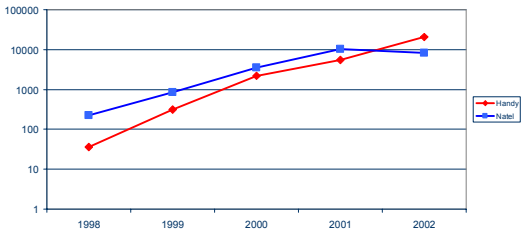
## Ergebnisse der AltaVista-Suche




| | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|
| Handy | 36 | 320 | 2233 | 5491 | 20309 |
| Natel | 224 | 850 | 3510 | 10240 | 8275 |



## Ergebnisse der AltaVista-Suche






## Das Web als Korpus

Vorteile

1. Grösse: > 1 Milliarde Dokumente
2. Verfügbarkeit: an jedem vernetzten Computer-Arbeitsplatz
3. Aktualität: ständig neue Dokumente

## Grösse des Web für eine Sprache

- Wie schätzt man die Anzahl der Wörter, die eine Suchmaschine für eine Sprache indexiert hat?
- Suche nach sehr häufigen Wörtern (also z.B. 'der', 'und'). Deren relative Häufigkeit pro 1000 Token ist (in etwa) konstant.



## Korpus-Zugriff

- auf ein lokal gespeichertes Korpus
  - via Datenbank-Schnittstelle oder
  - Spezialsoftware (z.B. Konkordanzprogramme)

  → vollständiges Durchsuchen möglich
- auf das Web
  - via Suchmaschine (wie Google oder AltaVista)

  → vollständiges Durchsuchen nicht möglich



## Suchmaschine

- liefert Fundstellen (geordnet nach Relevanz)
- liefert Häufigkeiten für die Suchbegriffe
- erlaubt
  - Einschränkung auf Sprache
  - Einschränkung auf Zeitraum
  - Einschränkung auf Domäne (z.B. nur Webadressen mit Suffix .ch)
  - Kombination von Suchbegriffen (AND, OR, NOT, NEAR)
  - Suche mit Mustervergleich (Wildcards: *,?)



## Die Nutzung des Web für die Linguistik

- Nutzen der gefundenen Dokumente
  - Lexikographie (Neologismen, Namen)
  - Parallele Texte für Übersetzer
- Nutzen der Häufigkeitsangaben
  - Übersetzung von Komposita
  - Disambiguierung auf Syntaxebene
  - Antwort-Validierung



## Extraktion und automatisiertes "Lernen" von Namen

- Suche nach Mustern:
  - "*die folgenden Politiker ...*"
    - Extraktion von Personennamen
  - "*Chemie-Firmen wie ...*"
    - Extraktion von Firmennamen
  - "*PCs der Serie ...*"
    - Extraktion von Produktbezeichnung



## Übersetzung von Komposita DE-EN
(Grefenstette 1999)

- *DE: Aktienkurs* → EN: ??
  - *Aktie* (*share, stock*) und
  - *Kurs* (*course, price, rate*)
- Erzeugung aller möglichen Übersetzungen (*share course, share price, share rate, stock course, ...*)
- Idee: Die Übersetzung mit der höchsten Häufigkeit im Web ist korrekt!

→ Führt zu 87% korrekten Entscheidungen.

## Parallele Texte für Übersetzer
(Resnik 1999: Mining the web for bilingual text)

- Wichtige Quelle für Übersetzer:
  - bereits bestehende Übersetzungen
- Problem: Finde übersetzte Texte im Web!
- Idee: Der Computer sucht nach Mustern wie:
  → *Click here for English version*
- Er lädt die Dokumente und vergleicht die Struktur (Länge, Anzahl Absätze, Zahlen, ...)



## Parallele Texte für Übersetzer
(Idee von M. Volk 2003 ;-)

Alternative zum Auffinden von parallelen Web-Dokumenten:

- Automatisches Übersetzen einer Webseite.
- Suchen nach "ähnlichen" Web-Seiten (in Google vorhanden) zu der Übersetzung (nicht vorhanden??).
- Vergleich der Struktur des Originals mit den gefundenen Dokumenten.



## Antwort-Validierung
(Magnini et al. 2002: Is it the right answer)

- Problem: Ein Frage-Antwort-System liefert eine Menge von möglichen Antworten zu einer gegebenen Frage.
- Idee: Wenn Frage(teile) und Antwort häufig im WWW gemeinsam vorkommen, dann ist es eine gute Antwort.
- Vorgehen: Extraktion von Schlüsselwörtern aus der Frage und der Antwort. Steigerung der Ausbeute via Muster-Kürzung.



## Antwort-Validierung (Beispiel)

- Frage: *Which river in the US is known as Big Muddy?*
- Antwort: *Mississippi River* and *Columbia River*
- Web-Suche 1: *[river* NEAR *US* NEAR *Big* NEAR *Muddy* NEAR *Mississippi* NEAR *River]*
- Web-Suche 2: *[river* NEAR *US* NEAR *Big* NEAR *Muddy* NEAR *Columbia* NEAR *River]*



## Synonym-Qualität
(Turney 2001: Mining the Web for Synonyms)

- Aus TOEFL gegeben
  - Ausgangswort: *levied*
  - Mögliche Synonyme: *imposed, believed, requested, correlated*
- Frage: Welches ist dem Ausgangswort in der Bedeutung am ähnlichsten?



## Synonym-Qualität

**Pointwise Mutual Information Score**

- `score(choice) = log (p(problem & choice) / (p(problem) * p(choice)))`

**can be simplified to**

- `score(choice) = (p(problem & choice) / p(choice))`

**Translates to**

- `score(choice) = hits(problem AND choice) / hits(choice)`

## Synonym-Qualität

Alternative

- `score(choice) = hits(problem NEAR choice) / hits(choice)`

**Oder**

- `score(choice) = hits((problem AND choice) AND NOT ((problem OR choice) NEAR "not")) / hits(choice AND NOT ((problem OR choice) NEAR "not") )`



## Synonym-Qualität

- Ergebnisse mt 80 TOEFL-Fragen

| | | |
|---|---|---|
| AND | 50/80 | 62,5% |
| NEAR | 58/80 | 72,5% |
| NEAR und NOT | 59/80 | 73.75% |
| College student | 51,6/80 | 64,5% |



## Disambiguierung von PP-Zuordnung

Das Problem: Mehrdeutigkeit

***Check deine Emails in der Badehose!***

- *Check*
  *(NP deine Emails (PP in der Badehose))*
- *Check* *(PP in der Badehose)*
  *(NP deine Emails)*



## Disambiguierung von PP-Zuordnung

Eine Lösung: Statistische Kookkurrenz
(Häufigkeit des gemeinsamen Auftretens im Verhältnis zum alleinigen Auftreten)

cooc(Verb, Präposition, PP-Nomen) =
freq(Verb, Präposition, PP-Nomen) / freq(Verb)



## Extraktion of 5-Tupeln

Satz: *Das Dorfmuseum gewährt nicht nur einen Einblick in den häuslichen Alltag.*

1. Verb: *gewährt*
2. Nomen1: *Einblick*
3. Prep.: *in*
4. Nomen2: *Alltag*
5. Funktion: postnominal modifier



## WWW Häufigkeiten

- Suchmaschinen liefern:
  **number of pages found**
  - freq(Noun1 NEAR Preposition NEAR Noun2)
  - freq(Noun1)
  - freq(Verb NEAR Preposition NEAR Noun2)
  - freq(Verb)

## Disambiguierung von PP-Zuordnung

Evaluation:

- Verfahren getestet über rund 4000 'manuell' disambiguierten Testfällen

Ergebnis der automatischen Disambiguierung:

- rund 2/3 der Testfälle sind entscheidbar
- korrekt entschieden: in 75% der Testfälle



## Alternative zur Nutzung der Suchmaschinen-Häufigkeiten

1. Web-Suche nach Dokumenten, die das Nomen N1 enthalten.
2. Herunterladen einer festen Anzahl der gefundenen Dokumente (z.B. 1000).
3. Extrahieren aller Sätze, die das gesuchte N1 enthalten.
4. Linguistische Verarbeitung dieser Sätze (PoS-Tagging, Lemmatisierung, Namen-Erkennung etc.)
5. Berechnen der Häufigkeiten über diesen Sätzen.



## Das Web als Korpus

Zusammenfassung

- Zugriff auf immer gleiche URLs (z.B. Projekt Wortwarte)
- Zugriff auf Häufigkeiten von Suchmaschinen
- Verarbeitung von Dokumenten, deren URLs durch Suchmaschinen geliefert werden.
- Aufbau einer eigenen Suchmaschine
- (Zugriff auf linguistische Ressourcen: z.B. LEO)



## Das Web als Korpus

Nachteile

- Grösse
- Aktualität: Ständig wechselnde Datenquellen
- (teilweise) unzuverlässige Datenquellen
- Nicht linguistisch strukturiert



## Flaschenhals: Suchmaschine

- ist auf inhaltliche Relevanz optimiert und nicht auf linguistische Fragen
- unterstützt keine linguistischen Operatoren
- erlaubt keine (genaue) Einschränkung auf Sachgebiete oder Textsorten



## Aktuelle Lösung: Linguistik-Filter

z.B. das Programm KWiCFinder

- erlaubt präzisere Anfrage
  - z.B. BEFORE, AFTER mit Abstand
- leitet Anfrage an AltaVista weiter
- bereitet Ergebnis für den Linguisten als Keyword-in-Context auf

## Zukunft: Eine Linguistik-Suchmaschine

- Linguistische Analyse **bei der Indexierung**
  - Lemmatisierung und Komposita-Segmentierung
    - Suche *Haus, Hauses, Häuser* und *Häusern*
  - Wortarten-Bestimmung
    - Suche dt. *'Junge'* als Nomen
    - Suche engl. *'can'* als Nomen
  - Phrasen-Erkennung
    - Suche *'mit + Kind'* in der selben Nominalphrase



## Zusammenfassung

- Korpora bieten neue Möglichkeiten für
  - Linguisten (insbes. Lexikographen)
  - Fremdsprachen-Lerner
  - Übersetzer
  - Computerlinguisten
- Das Web als Korpus
  - bietet grosse Chancen
  - bessere linguistische Erschliessung nötig



## How about ...

*He sees the man with the telescope.*

freq(*sees* NEAR *with*) = 244'865
freq(*sees*) = 1'806'082
→ cooc(*sees, with*) = 0.124

freq(*man* NEAR *with*) = 2'550'804
freq(*man*) = 14'444'376
→ cooc(*man,with*) = 0.176



## How about ...

*He sees the man with the telescope.*

freq(*sees* NEAR *with NEAR telescope*) = 150
freq(*sees*) = 1'806'082
→ cooc(*sees, with, telescope*) = $8.305 * 10^{-5}$

freq(*man* NEAR *with NEAR telescope*) = 478
freq(*man*) = 14'444'376
→ cooc(*man,with,telescope*) = $3.309 * 10^{-5}$