Webkorpora als qualitätsgesicherte Forschungsdaten

Marc Kupietz, Harald Lüngen, Cyril Belica, Rainer Perkuhn

Workshop:
Webkorpora in Computerlinguistik und Sprachforschung
Mannheim, 27.09.2012

Überblick

1. Einleitung
   - IDS/DEREKO-spezifische Bedingungen
2. Herausforderungen
   - Rechtliche Herausforderungen
   - Methodische Herausforderungen
3. Experiment
   - Gewinnung eines »Literaturkorpus« aus deWaC
   - Evaluation
4. Zusammenfassung

IDS/DeReKo-spezifische Bedingungen

- IDS: Stiftung des bürgerlichen Rechts
  - → Direktor haftet mit seinem Privatvermögen
- Satzungsauftrag: Wissenschaftliche Erforschung und Dokumentation der deutschen Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte.
- IDS ist im besonderen Maße auf einen guten Ruf Textgebern gegenüber angewiesen
  - da DeReKo und seine Nutzer von Textspenden abhängig sind

Besondere DEREKO-Eigenschaften

- very large general purpose corpus
- für wissenschaftliche Zwecke öffentlich zugänglich
- rechtlich unbedenklich
- aufgrund von Lizenzvereinbarungen nicht downloadbar, sondern nur über Analysesoftware zugänglich
- nicht nur für lexikografische, sondern allgemein für linguistische Anwendungen
- vor allem quantitative Anwendungen
- Urstichproben-Design: Nutzer müssen in der Lage sein, virtuelle Korpora u.a. anhand von Metadaten definieren zu können

»Web-Texte« z. Zt. im DeReKo-Archiv

- Wikipedia
- Wikipedia-Diskussionen
- Spektrumdirekt (Online-Magazin)
- Online-Ausgabe der Zeit
- nicht im Archiv:
  - deWaC Baroni/Kilgarriff (2006)
  - selbst gecrawlte Texte

Begriff »Web-Corpus« im Folgenden

- im Sinne von »Randomly Crawled Corpus«
- nicht: für das Medium Web typische Texte (Blogs, Foren, Chats, Tweets)
- nicht: Web nur als Weg der Zustellung (statt CD)

Urheberrecht

- Kopieren von Verbreiten von Web-Texten ist ebenso wenig erlaubt wie das Kopieren und Verbreiten von Buchtexten
- Schrankenregelung unzureichend für eine nachhaltige Verfügbarmachung
- Einholung von Lizenzen zu teuer
  - da Rechteinhaber zu zahlreich und schwer ermittelbar
- das IDS muss beim Vordringen in rechtliche Grauzonen zurückhaltend sein

→ z.Zt. und im Allgemeinen können Web-Corpora nicht über DeReKo persistent verfügbar gemacht werden

Ansätze zum Umgang mit der rechtlichen Problematik

- gezieltes Crawling von ausreichend lizensierten Dokumenten
  - aber: z.B. CC-Lizenzen leider selten ausgezeichnet
- Unterlaufen der Schöpfungshöhe durch Shuffeln von Sätzen
  - aber: für DEREKO keine Lösung und nicht ganz ohne Risiko
- zukünftig im Windschatten von Google?

Hoffnung: »Implied License«

- Konzept bekannt aus der US-amerikanischen Rechtsprechung:
- wer Inhalte online stellt, muss damit rechnen, dass diese den üblichen Gepflogenheiten entsprechend verwendet werden
- dem deutschen Rechtssystem eigentlich unbekannt
- aber: der Bundesgerichtshof hat in 2 Urteilen (2011, 2012) zugunsten von Google so eine implizierte Lizenz anerkannt und in seiner Begründung verwendet
- vielleicht ein Anfang, aber nichts, auf das das IDS bauen kann

Wenig Metawissen über aus dem Web gecrawlte Texte

- Analogie: Wir haben jede Menge Fußballergebnisse, wissen aber nicht, wer wann gegen wen gespielt hat
- (automatische) Übersetzung?
- Sprache, die wir nicht als deutsch bezeichnen würden?
- spezielle Web-Sprache?
- Zeit?
- . . .

Ist das wirklich so?

Typischerweise in herkömmlich akquirierten Texten vorhandene Metadaten:

- Datum der Veröffentlichung
- Herausgeber/Verlag
- Ort der Veröffentlichung
- Lizenzbedingungen
- Texttyp (Zeitungsartikel, Roman, Rede, Enzyklopädie-Artikel...)

Nicht sehr viele explizite Metadaten, aber . . .

- implizites Metawissen darüber, was für Texte in einem Korpus enthalten sind bzw. nicht enthalten sind
- explizite Metainformationen, bzw. potenziell relevante Strata lassen sich bei Bedarf daraus gewinnen
- z.B. (ad-hoc)-Textsorten:
  - Texte, die P. E. als Gebrauchstexte (o.a. »Gutes Deutsch«) definieren würde
- Unterschiede sind z.T. graduell aber gravierend

Offene Fragen

- In wie weit lassen Beobachtungen aus Web-Corpora Rückschlüsse auf eine Grundgesamtheit zu?
- Was ist die Grundgesamtheit/Sprachdomäne?
- Wie lässt sich das Web (zufällig/ systematisch/ stratifiziert) sampeln?
- Wie lassen sich virtuelle Korpora aus Web-Texten definieren?

Ansätze das Web trotzdem für DEREKO nutzen

1. Web-Crawling vor allem zur Identifikation interessanter neuer Textgeber, deren Texte dann auf herkömmliche Weise akquiriert werden.
2. Vergleich von DEREKO mit Web-Corpora um unterrepräsentierte Strata zu identifizieren
3. Texte gezielt, z.B. im Hinblick auf unterrepräsentierte Strata crawlen, etwa durch gezielte Auswahl von Seeds, mit Einschränkung auf CC-Lizenz
4. Versuchen, Meta-Informationen zu einzelnen Texten automatisch zu ermitteln.
5. Versuchen, interessante Teilkorpora aus Web-Korpus-Archiven zu gewinnen

Gewinnung interessanter Teilkorpora aus Web-Corpora

Grundidee:

- Crawling großer Web-Korpusarchive sehr preiswert
  - auch dann noch, wenn ein Großteil weggeworfen wird
- die Eigenschaften eines interessanten/ wertvollen Korpus können durch ein Beispielkorpus definiert werden

Vorgehensweise:

1. Definition der Eigenschaften des Ziel-Korpus durch Erstellung eines entsprechenden Referenz-Korpus RC aus bekannten Korpusdaten
2. Festlegung eines Distanzmaßes D, das die (Un-)Ähnlichkeit zweier Korpora angibt
3. nimm das Web-Corpus-Archiv WC, und streiche daraus Texte, so dass $D(RC, WC')$ minimiert wird

Ähnlichkeitsmaß für den Vergleich von Korpora

- Vorschlag von Kilgarriff (2001):
  Distanzmaß basierend auf Wort-Frequenzlisten
- Ergebins seiner empirischen Ermittlung anhand von BNC-Teilkorpora bekannter Ähnlichkeit (bzgl. »text type«):
  - gute Datengrundlage: die 500 häufigsten Tokens der Vereinigung der beiden Korpora
  - χ^2-basiertes Distanzmaß am erfolgreichsten
  - 2.-bestes Distanzmaß: Spearmans Rangkorrelationskoeffizient
    - Vorteil: unabhängig von Korpusgrößen
    - Nachteil: Signifikanz von Rangunterschieden hochfrequenter Wörter wird evtl. unterbewertet

1. Einleitung
   - IDS/DeReKo-spezifische Bedingungen
2. Herausforderungen
   - Rechtliche Herausforderungen
   - Methodische Herausforderungen
3. Experiment
   - Gewinnung eines »Literaturkorpus« aus deWaC
   - Evaluation
4. Zusammenfassung

Ausgangsbasis

- virtuelles Referenzkorpus »Literatur des 20. und 21. Jh.« (»lit«) aus DEREKO (7.1 Mio. Wörter)
  - wertvoll, da Akquisition teuer und Stratum in DEREKO schwach besetzt
- WC-Archiv: deWaC-2006 (»dwc«): 1.4 Mrd. Wörter
- Distanzmaß $D_{\chi^2,500} = \frac{\chi^2}{500}$
- Ausgangsdistanz zwischen lit und einer gleich großen Stichprobe aus dwc: 1982.35
- zum Vergleich: Distanz zwischen taz10 und taz11: 17.12 (typischer Wert für aufeinanderfolgende Zeitungsjahrgänge)

Erzeugung von litdwc aus dwc

TA-Algorithmus, hier vereinfacht dargestellt

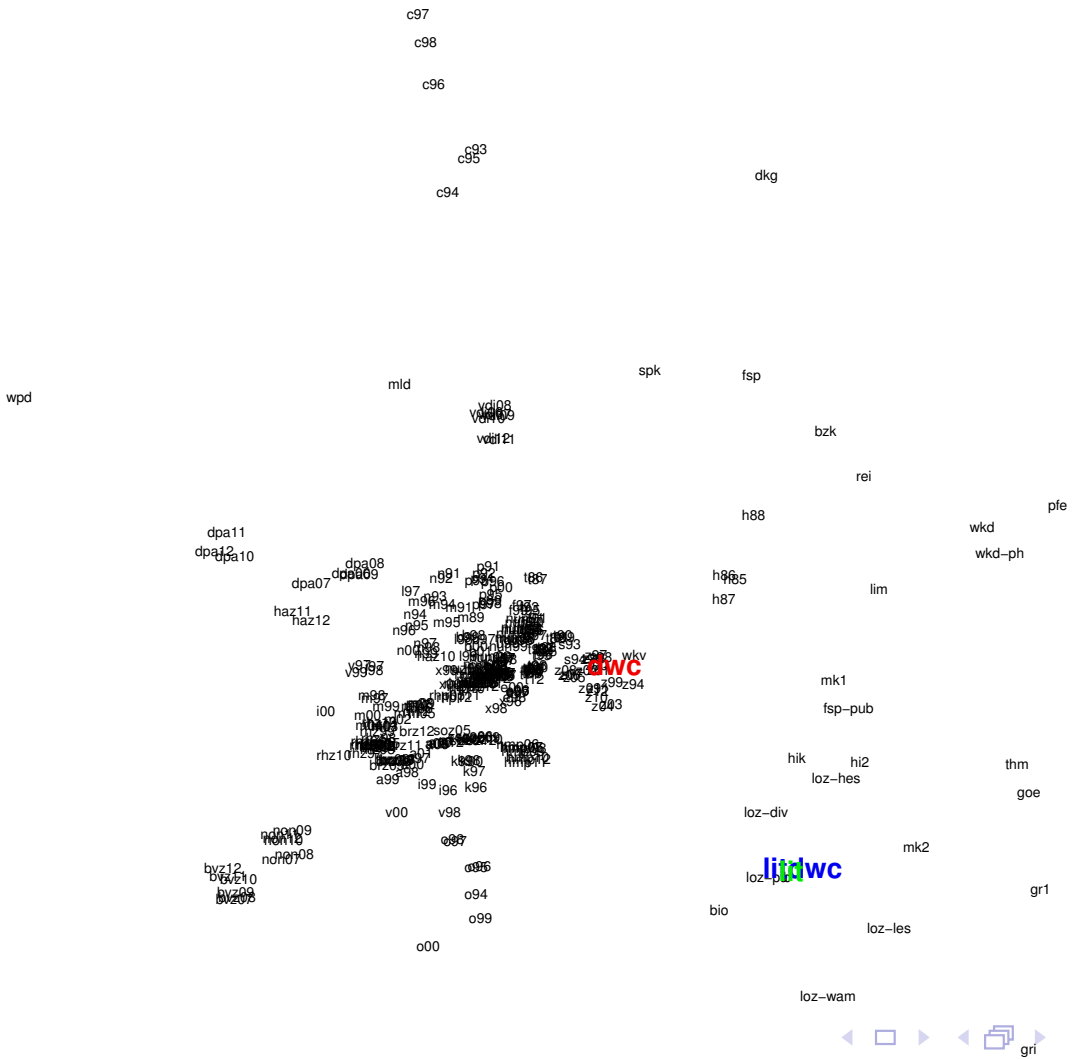
1. nimm lit als Referenz
2. nimm dwc als Basis
3. gehe alle (noch übrigen) Texte in dwc′ zufälligerer Reihenfolge durch und entferne Text, wenn dadurch D_{χ^2} kleiner wird
4. gehe alle entfernten Texte durch und füge Text hinzu, wenn dadurch D_{χ^2} kleiner wird
5. breche ab, wenn über mehrere Zyklen keine Verbesserung, sonst weiter mit 3.

Ergebnis

- resultierendes Korpus litdwc: 8.2 Mio. Wörter
- $D_{\chi^2}(lit, litdwc) = 9.28$
- Ausgangsdistanz: $D_{\chi^2}(lit, dwc_s) = 1982.35$
- $D_{\chi^2}(lit, litdwc)$ ist kleiner als die Distanz zwischen den meisten aufeinanderfolgenden Zeitungsjahrgängen

MDS-Abb. der ρ-Distanzmatrix aller DEREKO-Subkorpora

mit dwc, lit und litdwc (hier Spearmans Rangkorrelationsk., nicht χ^2!)



Diskussion

Es konnte aus deWaC ein Korpus gewonnen werden, das dem Literatur-Referenzkorpus bzgl. D_{χ^2} sehr ähnlich ist.

Offene Fragen:

1. Müssten für das Deutsche nicht mehr als die 500 häufigsten Tokens betrachtet werden?
2. Wie robust ist D_{χ^2} und die verwendete Methode?
3. Heißt ein kleines D_{χ^2} wirklich, dass die »Sprache« ähnlich ist
   - oder sind nur isolierte lexikalische Eigenschaften (z.B. Verwendung von Vornamen)
   - oder linguistisch randständige Aspekte (z.B. Thema) ähnlich?

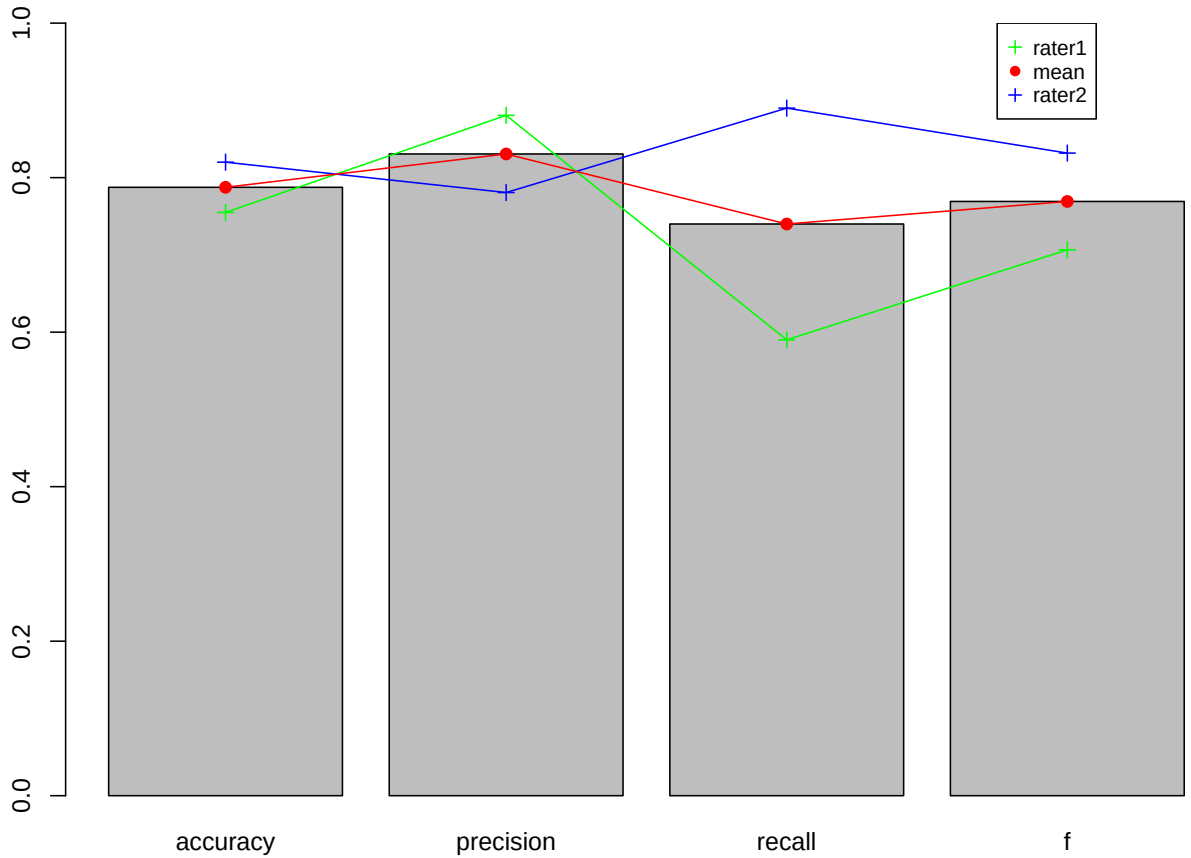
Evaluationsstudie

- zufällige Auswahl von je 100 Sätzen aus
  - lit
  - litdwc
  - litdwc$_{lc}$[1]
  - dwc\litdwc
  - dereko\lit
- 2 Rater/VPs (SHKs, Muttersprachler) folgendermaßen instruiert: »Markieren Sie die Sätze, die Ihrer Meinung nach aus einem belletristischen Werk (Roman, etc.) kommen, mit 1 und die, die nicht aus einem belletristischen Werk kommen, mit 0.«

[1] wie litdwc gewonnen, aber nur klein geschriebene Wörter berücksichtigt

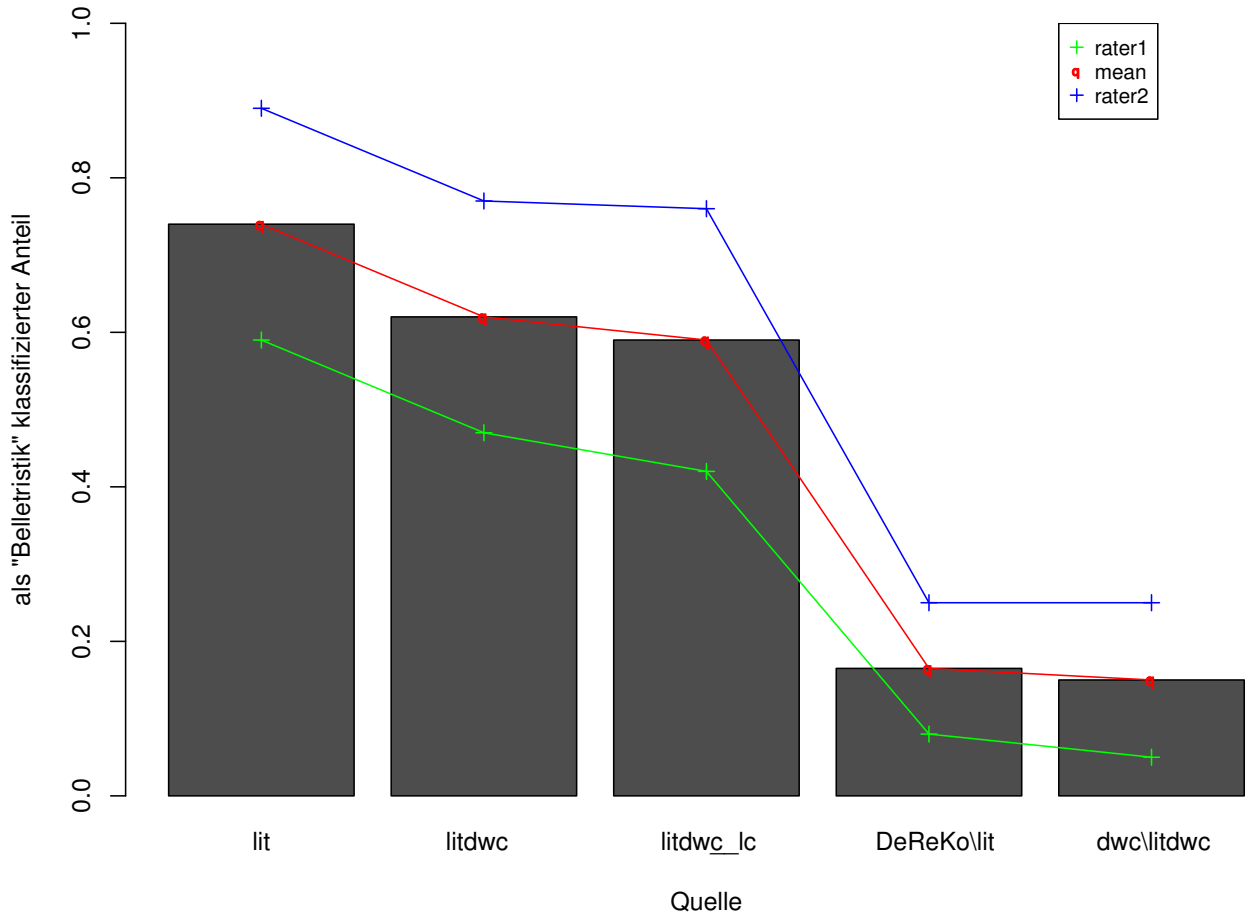
Evaluation der Einschätzungen

basierend auf den Sätzen aus lit und DeReKo\lit



agreement 0.702
kappa 0.438

Ergebnisse der Evaluationsstudie



Von beiden Ratern als Belletristik eingeschätzt

- Die Schlange Ka war ihm aus längst versunkenen Tagen noch lebhaft in Erinnerung.
- Nach etwa einer Stunde waren wir an einem Fluß angelangt, an dem der Lkw entlangfuhr.
- Wie warm ihre Hand war !
- In mir pulsierte ein neues, ein stärkeres Herz, eine Kraft, die alles in den Schatten stellte, was ich bisher gekannt hatte.
- Stell dir vor, da oben steht ein Selbstmordattentäter und stürzt sich auf uns runter, sagt einer der Freunde.
- Johnny würde einen Platz am Rand einnehmen, alsbald Mittelpunkt, und, ohne sich von der Stelle zu bewegen, jedes Geschehen kommentieren, zu dem man ihm das Stichwort gab.
- Sie war konzentriert und ganz außerhalb der Realität gewesen und musste sich ihrer Umgebung erst wieder bewusst werden.

Unterschiedlich eingeschätzte Sätze

- Dort können sie sich tage-, ja wochenlang aufhalten.
- Also, bis dann...
- Aus Erleichterung habe ich mich zu fest an den Stehtisch gelehnt, und er brach krachend zusammen.
- SO KOMMT ES ALSO NICHT AUF DEN WOLLENDEN ODER LAUFENDEN AN, SONDERN AUF DEN SICH ERBARMENDEN GOTT ?
- Oder ist es schon nach Sibirien gewandert ?
- Der zweite Zapfhahn wurde herausgeklappt, der dritte, plötzlich standen vier Leute hinter der Theke, dann fünf, und zwei wurden wieder nach vorne bugsiert - Freibiergesichter.
- Im wahrsten Sinne des Wortes.
- Na dann lassen wir uns eben überraschen, und nun laß Mami weiter arbeiten.

Von beiden Ratern als Nicht-Belletristik eingeschätzt

- Seine Beschäftigung mit der Steuerung von Geschützen während des zweiten Weltkriegs führte ihn zu der Weiterentwicklung der Nachrichtentechnik zur Kybernetik.
- Briefe haben wir mit 10 Pfennig-Marken frankiert, der Postbote trug sie mindestens 2 x am Tag aus.
- Unsere Partnerschaft besteht seit März 1993.
- Inzwischen sind wir klüger.
- In Nordrhein-Westfalen sei sogar ein ' Bündnis für Erziehung ' ins Leben gerufen worden, das ebenfalls den Willen der Verantwortlichen verdeutliche, den ihnen obliegenden erzieherischen Verpflichtungen gerecht zu werden.
- Es war eine jener Veranstaltungen, bei der DDR-Bürger mit Westdeutschen, aber auch DDR-Bürger untereinander, die sich sonst kaum sahen, zusammen kamen.
- Ran an die Tasten;-) - > Jetzt kennenlernen !

Diskussion

- die Sätze aus beiden erzeugten »Literaturkorpora« wurden deutlich häufiger als »Belletristik« eingeschätzt als die aus den Baseline-Kontrollkorpora

→ die gewählte Methode zur Generierung von Subkorpora aus deWaC war diesbezüglich erfolgreich

→ $D_{\chi^2,500}$ erfasst Aspekte, die für die Einschätzung von Sätzen als »Belletristik« relevant sind

- welches genau diese Aspekte sind und welche Aspekte nicht erfasst werden, muss noch genauer untersucht werden
- die URLs von litdwc und litdwc$_{lc}$ können unter `http://corpora.ids-mannheim.de/litdwc` heruntergeladen werden

Zusammenfassung

- im IDS/DEREKO-Kontext gibt es besondere Herausforderungen bzgl. der Verwendung von Web-Corpora
- die rechtlichen Voraussetzungen für eine Verfügbarmachung von zufällig gecrawlten WC sind z.Zt. nicht gegeben
- trotzdem kann das Web als Quelle in den Ausbau von DEREKO einbezogen werden
- die vorgestellte Methode zur Generierung von Subkorpora könnte einen Beitrag zur Lösung methodischer Probleme mit WC liefern

Vielen Dank!

korpuslinguistik@ids-mannheim.de

Referenzen

Baroni, M. / Kilgarriff, A. (2006): Large linguistically-processed web corpora for multiple languages. In: *Conference Companion of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, 87–90.

Kilgarriff, A. (2001): Comparing corpora. *International Journal of Corpus Linguistics*, 6 (1), 97–133. `http://www.kilgarriff.co.uk/Publications/2001-K-CompCorpIJCL.pdf`.

Kupietz, M. / Belica, C. / Keibel, H. / Witt, A. (2010): The German Reference Corpus DEREKO: A primordial sample for linguistic research. In: *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC'10)*. European Language Resources Association (ELRA).