The Good, the Bad, and the Hazy: Design decisions in web corpus construction

Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer
German Grammar and Linguistics (FU Berlin)

WaC8, Lancaster, July 22, 2013

COW project:
http://hpsg.fu-berlin.de/cow/

texrex (current version: texrex-hyperhyper):
http://sourceforge.net/projects/texrex/

Our brand new book on web corpus construction:
http://dx.doi.org/10.2200/S00508ED1V01Y201305HLT022
http://sites.morganclaypool.com/wcc/

Overview

Text quality

Rating experiment

Badness scores

Design decisions and non-destructive normalization

Freie Universität Berlin

We are here. . .

Text quality

Rating experiment

Badness scores

Design decisions and non-destructive normalization

Text quality as understood here

- sentences
- ideally connected
- **not** word/name lists
- **not** tag clouds

Freie Universität Berlin

The Good

Search the SEP

• Advanced Search
• Tools • Random Entry

Table of Contents
• What's New
• Archives
• Projected Contents

Editorial Information
• About the SEP
• Editorial Board
• How to Cite the SEP
• Special Characters

Support the SEP
• PDFs for SEP Friends
• Make a Donation
• SEPIA for Libraries

Contact the SEP

© Metaphysics Research Lab, CSLI, Stanford University

1. Introduction

Both logic and ontology are important areas of philosophy covering large, diverse, and active research projects. These two areas overlap from time to time and problems or questions arise that concern both. This survey article is intended to discuss some of these areas of overlap. In particular, there is no single philosophical problem of the intersection of logic and ontology. This is partly so because the philosophical disciplines of logic and of ontology are themselves quite diverse and there is thus the possibility of many points of intersection. In the following we will first distinguish different philosophical projects that are covered under the terms 'logic' and 'ontology'. We will then discuss a selection of problems that arise in the different areas of contact.

'Logic' and 'ontology' are big words in philosophy, and different philosophers have used them in different ways. Depending on what these philosophers mean by these words, and, of course, depending on the philosopher's views, sometimes there are striking claims to be found in the philosophical literature about their relationship. But when Hegel, for example, uses 'logic', or better 'Logik', he means something quite different than what is meant by the word in much of the contemporary philosophical scene. We will not be able to survey the history of the different conceptions of logic, or of ontology. Instead we will look at areas of overlap that are presently actively debated.

2. Logic

There are several quite different topics put under the heading of 'logic' in contemporary philosophy, and it is controversial how they relate to each other.

2.1. Different conceptions of logic

On the one hand, logic is the study of certain mathematical properties of artificial, formal languages. It is concerned with such languages as the first or second order predicate calculus, modal logics, the lambda

The Bad

Home
Blog
Luftfracht Speditionen
Logistik Videoportal
Speditionen
Umzugsspeditionen Deutschland
Logistikzielorte in Deutschland

LOGISTIKJOURNALE

- Cargo Journal
- Umzug Journal
- Internationale Speditionen

LOGISTIK ZIELORTE IN DEUTSCHLAND

- Logistik Zielorte in Baden Württemberg
- Logistik Zielorte in Bayern
- Logistik Zielorte in Berlin
- Logistik Zielorte in Brandenburg
- Logistik Zielorte in Bremen
- Logistik Zielorte in Hamburg
- Logistik Zielorte in Hessen
- Logistik Zielorte in Mecklenburg-Vorpommern
- Logistik Zielorte in Niedersachsen
- Logistik Zielorte in Nordrhein-Westfalen

Liste deutscher Speditionen

Liste internationaler Speditionsunternehmen

Amm Spedition	Amm GmbH & Co KG Spedition
Amm Familie	90451 Nürnberg
Anhalt Logistics	Anhalt Logistics GmbH & Co. KG
Anhalt Familie	25776 Rehm-Flehde-Bargen
Arriva	arriva gmbh
	79115 Freiburg
Beck Gruppe	Beck Spedition+Logistik GmbH
Beck/Kienzler	70794 Filderstadt
Biber Post	Marketing Service Magdeburg GmbH
	39104 Magdeburg
BTG Feldberg	BTG Feldberg & Sohn GmbH & Co. KG
Feldberg Familie	46395 Bocholt
BurSped Gruppe	KG Bursped Speditions-GmbH & Co

Freie Universität Berlin

The Hazy

Zutaten für 4 **Portionen** umrechnen

100 g **Marzipan - Rohmasse**
300 ml Milch
3 **Ei(er)**
7 EL Mehl
Für die Füllung:
125 g Mohn - Mischung, backfertig
Butterschmalz zum Ausbacken

Zubereitung

Marzipanrohmasse mit 2 EL Milch geschmeidig rühren. (Am besten mit einem Blitzhacker)
Eier mit Marzipanrohmasse mit dem Schneebesen des Handrührgerätes verquirlen, Die restliche Milch und das Mehl zugeben. Alles zu einem glatten Teig verrühren. 10 Minuten quellen lassen.
Das Butterschmalz in einer Pfanne erhitzen und aus dem Teig darin nacheinander vier goldgelbe Crêpes ausbacken. Die fertig gebackenen Crêpes warm halten.
Die Marzipancrêpes mit der Mohnmasse füllen, zu Dreiecken zusammenfalten.

Arbeitszeit:	ca. 20 Min.
Schwierigkeitsgrad:	normal
Brennwert p. P.:	keine Angabe
Freischaltung:	07.09.06
Rezept-Statistiken:	12.081 (156)* gelesen 148 (0)* gespeichert 439 (5)* gedruckt 14 (0)* verschickt * nur in diesem Monat

Verfasser:
feuermohn
Mitglied seit 09.12.2004
10.466 Beiträge (ø3,68/Tag)

Festliches Eisvergnügen mit wenig Aufwand und großer Wirkung

Schlagworte für dieses Rezept

☐ Dessert, ☐ Mehlspeisen, ☐ Süßspeise

Kombi-Suche

Ähnliche Rezepte

- Semlor
- Pfaffenhüetli-Zitronen
- Spekulatius, gefüllt
- Mandelhippen
- Schoko - Marzipan - Herzen
- Marzipan - Pistazien - Creme
- Pistazieneisparfait mit Nougatsauce
- Schoko - Marzipan - Eis
- Zwetschgennudeln
- Scheiterhaufen mit Äpfeln und Marzipan

Rezeptsammlungen

Dieses Rezept ist in diesen Sammlungen gespeichert:

- Sweeties
- Marzipan
- Kuchen
- Mehlspeisen als Hauptgericht
- Dessert

Freie Universität Berlin

The Hazier

Startseite EUROPAGES Geschäftsverzeichnis > Alle Geschäftsbereiche > Gummi und Rohstoffe > Kunststofferzeugnisse für das Baugewerbe

3580 Unternehmen für: **Kunststofferzeugnisse für das Baugewerbe**

Die folgende Liste enthält alle Lieferanten, Hersteller und Händler, die Ihrer Suche nach **Kunststofferzeugnisse für das Baugewerbe in der Branche Gummi und Rohstoffe** entsprechen.

Die auf dieser Seite aufgeführten Unternehmen passen auch zu folgenden Schlüsselbegriffen: kunststoffe, pvc-brunnen, pvc-fittings, baustoffe, pvc-rohre.

Wählen Sie mehrere Unternehmen aus und **Kontakt**

FIESSO UMBERTIANO (RO) - ITALIEN

LARETER SPA

Das Unternehmen LARETER arbeitet seit 1961 in der Kunststoffverarbeitung. Dank seines Know-hows und seines hohen Spezialisierungsgrads genießt das Unternehmen weltweites Renommé (Export in 27 Länder)...

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | fittings pvc artesische brunnen | einleitungen | polyethylenanschlüsse für bewässerungssysteme | bewässerung | pvc hochbau | rohrverbindungsstücke (fittings) aus kunststoff | gasleitungen | plastikanschlussstücke | gummiverbindungsstücke | pvc ...

http://www.lareter.it

Cholet Cedex - FRANKREICH

NICOLL RACCORDS PLASTIQUES

„Führender europäischer Hersteller von Produkten aus Synthesematerialien für den Hoch- und Tiefbau. Als Spezialist für Einspritzung und Strangpressen bietet Nicoll eine 3 Hauptbereiche umfassende...

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | kunststofffittings | rückschlagventile | unterbauten | lüftungsgitter | hydraulische abflussrinnen ...

http://www.nicoll.fr

Sainte Sigolène Cedex - FRANKREICH

GROUPE BARBIER

Die Gruppe BARBIER ist seit 50 Jahren auf die Herstellung von Kunststofffolien für die Landwirtschaft, Industrie und den Vertrieb von Kunststoffsäcken spezialisiert. Extrusion, Druck, PE-Schweißen....

Lieferant für: Kunststofferzeugnisse für das Baugewerbe | bedruckte folie | kunststofffolie | stretchfolie | industriefolien | kunststofffolien für die landwirtschaft | silage | kunststoffverarbeitung ...

http://www.barbiergroup.com/

Freie Universität Berlin

We are here. . .

Text quality

Rating experiment

Badness scores

Design decisions and non-destructive normalization

Corpus

- UKCOW2012 (beta version) [Schäfer and Bildhauer, 2012]
- approx. 6 GT, crawled in 2012 from `.uk`
- cleaned with `texrex-mrvain`
  - HTML stripping, conversion to ISO-8859, other normalizations
  - boilerplate removal
  - aggressive w-shingling-based deduplication
- evaluated as superior to ukWaC in collocation extraction tasks in Biemann et al. [2013] (or at least "equally good, but bigger")

Task for human raters

- classification of text quality for 1,000 documents
- 500 documents from the early phase of the crawl: **"early data"**
- 500 documents from the late phase: **"late data"**
- boilerplate removed, presented as text-only with paragraph breaks
- scale:
  - $-2, -1$: document **should not** be in the corpus
  - $1, 2$: document **should** be in the corpus
  - 0: undecided/document **might or might not** be in the corpus

From the guidelines I

- Documents containing predominantly full sentences are good, "predominantly" meaning considerably more than 50% of the text mass (as perceived by the coder).
- Boilerplate material in sentence form is good (*You are not allowed to post comments in this forum.*), other boilerplate material is bad (*Copyright © 2046 UAC Ltd.*).
- Sentences truncated or otherwise destroyed by some post-processing method are good as long as they are recognizable as (the rest of) a sentence.
- Repetitions of good sentences are good.

From the guidelines II

- Decisions should not depend on the length of the document, such that a document containing only one good sentence would still be maximally good.
- Non-English material contributes to badness.
- Non-sentence material (lists, tables, tag clouds) contributes to badness.
- However, if a list etc. is embedded in a coherent text which dominates the document, the document is good (prototypically recipes with longer instructions).

Raters

- raters **A** and **R**: corpus designers
  with a shared understanding of desired corpus
- rater **S**: student assistant,
  experience with three similar rating tasks
- "training": rating of 100 documents together
  with several hours of discussion of borderline cases

Hazy results

statistic	early 500	late 500	all 1,000
raw	0.566	0.300	**0.433**
κ (raw)	0.397	0.303	0.367
$ICC(C,1)$	0.756	0.679	**0.725**
raw ($r \geqslant 0$)	0.900	0.762	**0.831**
raw ($r \geqslant 1$)	0.820	0.674	0.747
κ ($r \geqslant 0$)	0.673	0.625	**0.660**
κ ($r \geqslant 1$)	0.585	0.555	0.598
κ ($r \geqslant 2$)	0.546	0.354	0.498

Acceptable results?

- **values below** 0.68 **[Krippendorff, 1980]**
- even considering criticism of Krippendorf's magic number [Carletta, 1996, Bayerl and Paul, 2011]:
  uncomfortably low for "gold standard"
- more confusion on late data (lower overall quality)
- worse: disagreement between corpus designers
- acceptance at the threshold $\geqslant 0$:
  A: 78.4%, R: 73.8%, S: 84.9%

Freie Universität Berlin

We are here. . .

Text quality

Rating experiment

Badness scores

Design decisions and non-destructive normalization

General method and idea

- simple metric with known properties
- language-independent, unsupervised...
  does not involve an obviously difficult design decision
- strategy for cleansing: **high recall for everyone**,
  accept mediocre precision
- for retained documents: use as annotation in final corpus,
  allow corpus users to "set" precision

Implementation

- based on "frequent/short word" method
  in language identification [Grefenstette, 1995]
- similar to WaCky [Baroni et al., 2009]
  but without manually compiled lists of function words
- totally unsupervised procedure for crawled data
  predominantly in a single language (TLD crawl):
  - **training**: get weighted mean and standard deviation
    of relative frequencies of the most frequent words ("profile")
  - **production**: calculate for the top m of them
    the "standardized" negative deviation for each document
  - clamped and added up: the **Badness score**

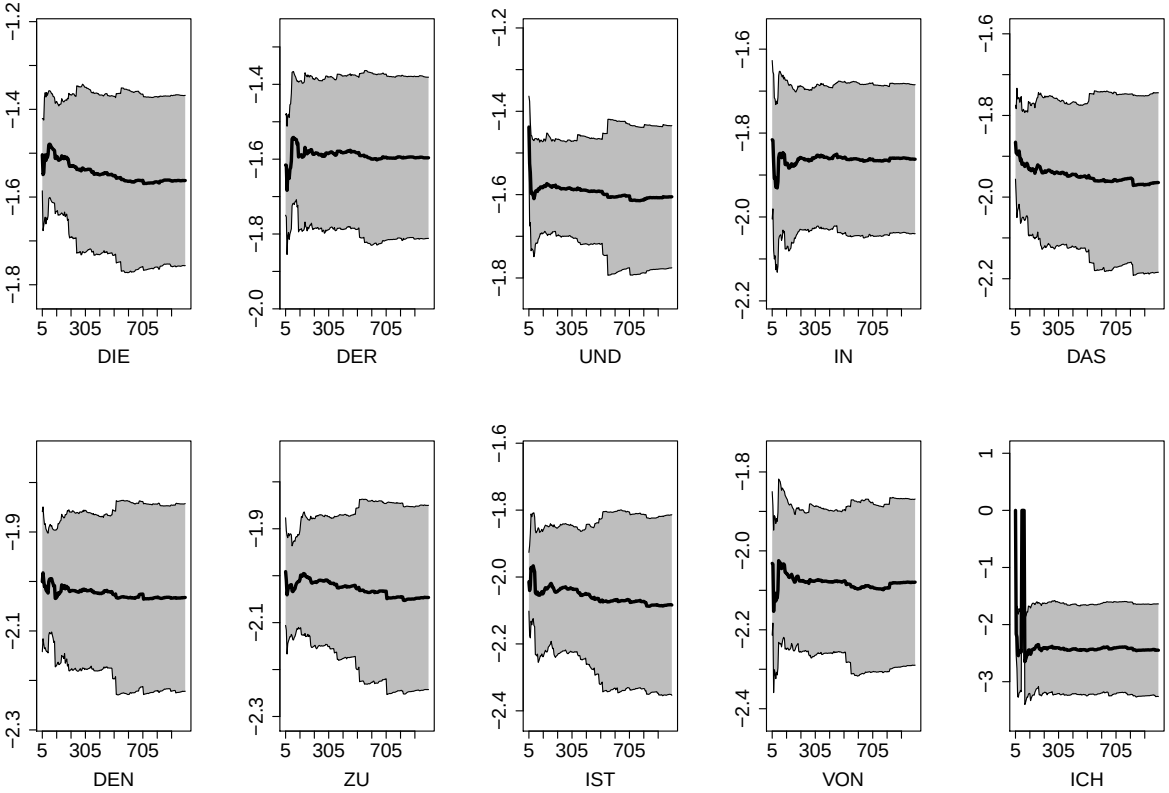
Questions

Does **profile generation**
yield reliable results?

How should we select the **threshold**
for the actual deletion of documents?

Freie Universität Berlin

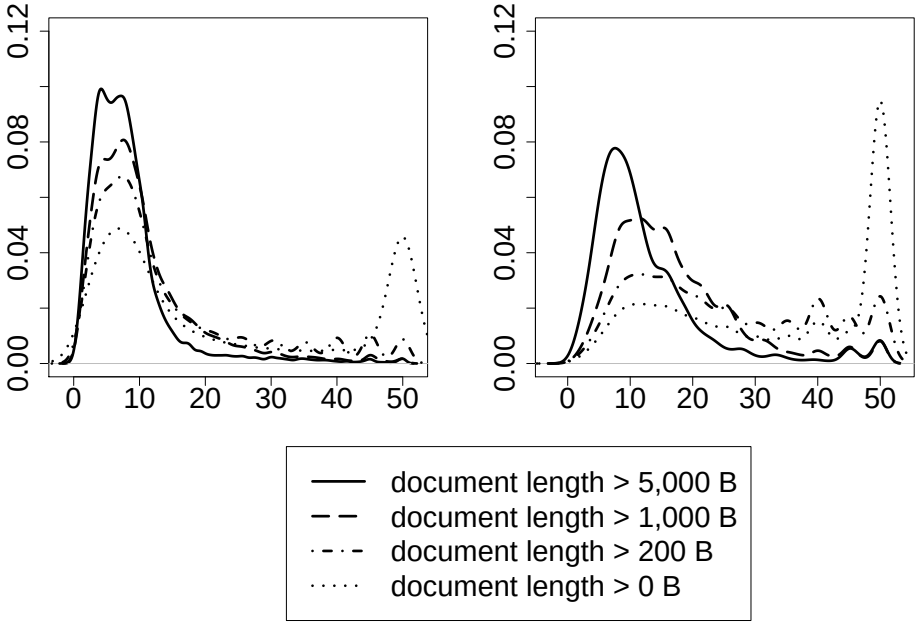
Profile development while training



German test profile; $n = 1000$; log10-transformed relative frequencies

Freie Universität Berlin

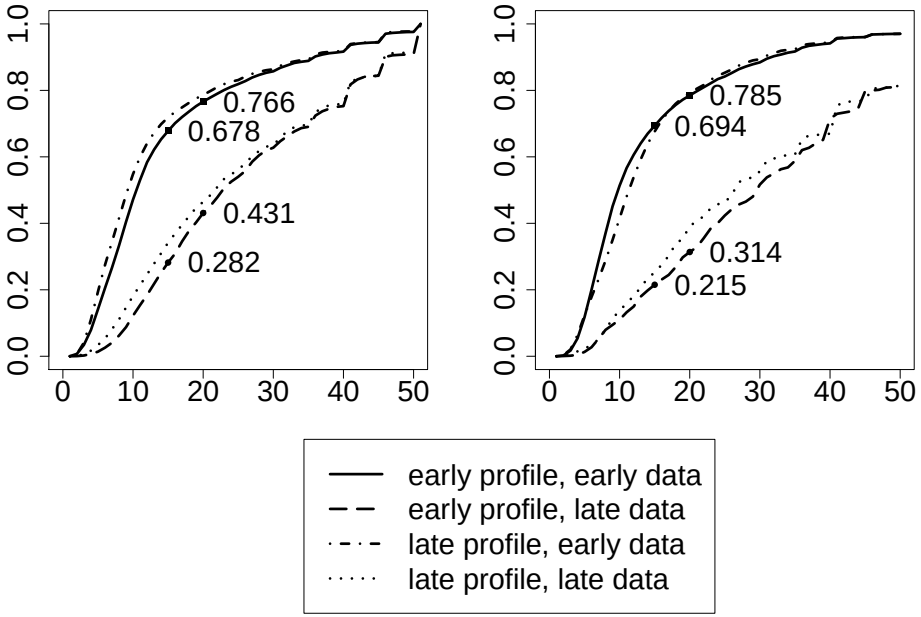
Distribution of Badness depending on document length (early profile on early data)



left: DECOW2012; right: UKCOW2012

Freie Universität Berlin

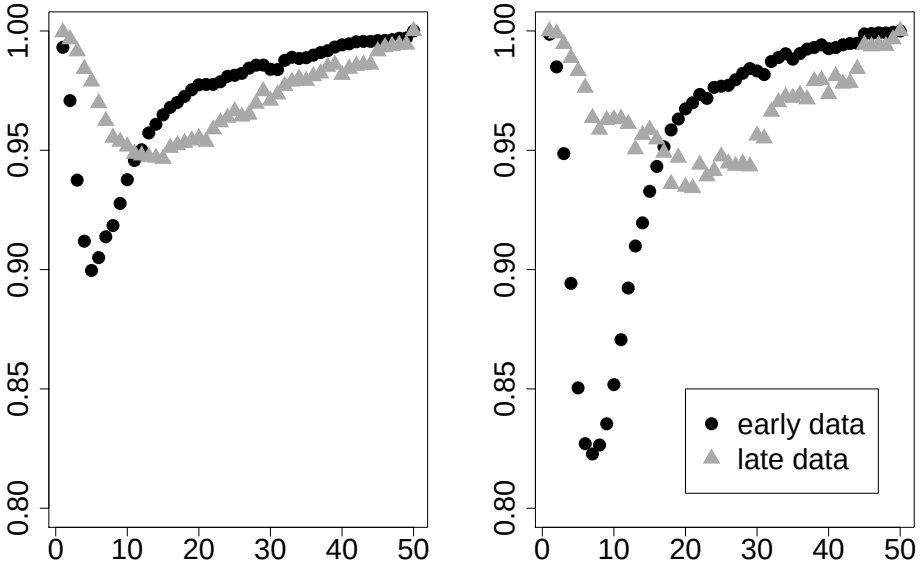
Profile comparison:
Effect of thresholds (=cumulative density of Badness)



left: DECOW2012; right: UKCOW2012; only documents over 200 B;
values at Badness 15 and 20 marked

Freie Universität Berlin

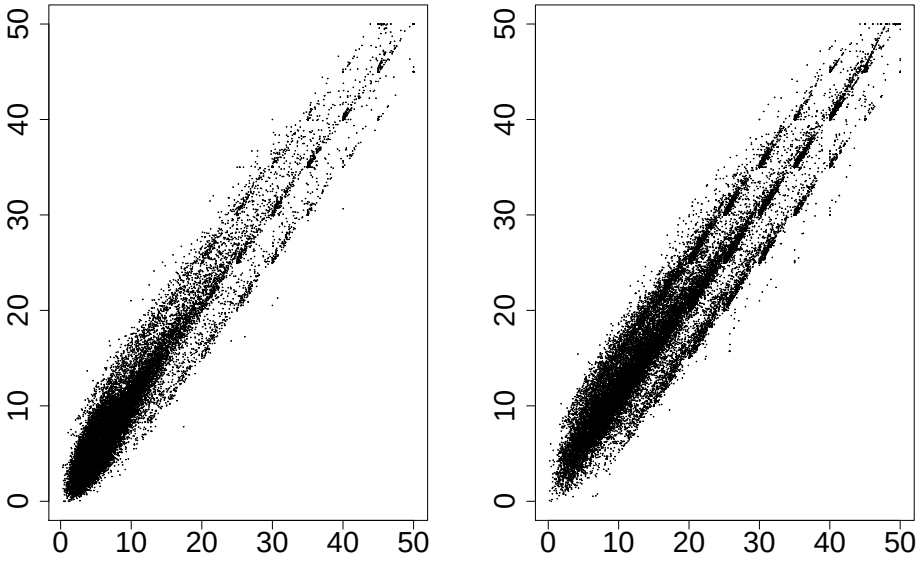
Profile comparison:
Raw agreement between profiles



left: DECOW2012; right: UKCOW2012

Freie Universität Berlin

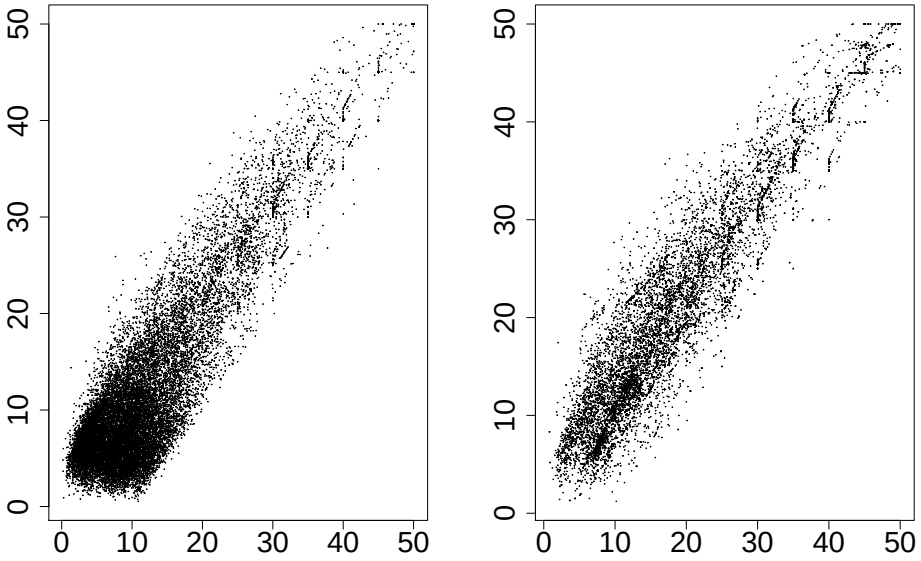
DECOW2012 sample raw profile comparison



left: early data; right: late data; x-axis: late profile; y-axis: early profile

Freie Universität Berlin

UKCOW2012 sample raw profile comparison



left: early data; right: late data; x-axis: late profile; y-axis: early profile

We are here. . .

Text quality

Rating experiment

Badness scores

Design decisions and non-destructive normalization

Badness threshold for removal

If recall is more important than precision: 35

	prec	rec	F1	correct	baseline
S	0.914	**0.959**	0.936	0.888	0.849
A	0.856	**0.973**	0.911	0.851	0.781
R	0.808	**0.976**	0.884	0.811	0.738

Precision/recall etc. for our three raters (documents better than 0)
at a Badness threshold of 35

Reasons for non-destructive normalization

- our goal: carefully sampled and processed web corpora for fundamental research – theoretical linguistics, linguistic web characterization
- **noise or distortion through processing intolerable**
- **Leave major destructive design decisions to the user!**

Freie Universität Berlin

Annotation with quality metrics instead of removal: A preliminary version based on UKCOW2012

```
<doc url="http://www.booktrust.org.uk/writing/writing-tips/33" bdc="d">
<p bpc="c">Accessibility and text options</p>
<p bpc="e">Professionals</p>
<p bpc="c">How to write comedy</p>
<p bpc="a">You want to write comedy?  Well it's easy.  I've just done it and didn't
even use a spell check.  I'd definitely suggest starting with a 'c' rather than a 'k'
as that sounds either wacky or German and either way there will be prejudices about
what type of humour it is.  Oh, sorry, comedy writing?  I see.  Well first try not
doing the most God-awful joke you can in your very first sentence.  That's a definite
no-no.  To be honest, I can't tell you how to write comedy as such.  I'm very much a
subscriber to the 'either you're funny or you're not' party.  However even if you have
the invite to that very party, there's a high chance you might not know quite how to
utilise the correct chat for the kitchen, or the right outfit to wear, in the same way
I've never worked out how to do a decent analogy.</p>
<p bpc="a">What I'm saying is that if you're a funny chap or chapette, there are some
things that definitely hone those powers into a neat bit of witty prose, rather than a
bundle of nearly funny cons.</p>
```

Annotation for **Badness**, **boilerplate**

Bottom line

- "Web corpus cleansing" is a destructive process.
- Even corpus designers achieve only mediocre agreement w. r. t. the "quality corpus material"/"noise" decision.
- Better guidelines just force corpus users to live with more specific problematic destructive decisions.
- Strategy: Leave as much as possible in the corpus and provide documented and intuitive quality annotation.
- Badness is such a metric.

References I

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3): 209–226, 2009.

P. S. Bayerl and K. I. Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, 2011.

C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch. Scalable construction of high-quality web corpora. *Special issue of JLCL*, 2013. In prep. The list of authors is preliminary and might reflect neither the order nor the actual list in the printed version.

J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

G. Grefenstette. Comparing two language identification schemes. In *Proceedings of the 3rd Internation conference on Statistical Analysis of Textual Data (JADT 1995)*, pages 263–268, Rome, 1995.

K. Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, 1980.

R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, 2012. ELRA.