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
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
The Good

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

Liste deutscher Speditionen

Liste internationaler Speditionsunternehmen

Umzug123

100% Kostenlos & unverbindlich

Umzugsfirmen finden und vergleichen.

Amm Spedition
Amm Familie
Anhalt Logistics
Anhalt Familie
Arriva

Beck Gruppe
Beck-Kienzle
Biber Post

BTG Feldberg
Feldberg Familie
Burspeed Gruppe

Amm GmbH & Co KG Spedition
90451 Nürnberg
Anhalt Logistics GmbH & Co KG
25776 Rehn-Fehnde-Bargen
Arriva gmbh
79115 Freiburg
Beck Spedition Logistik GmbH
70794 Filderstadt
Marketing Service Magdeburg GmbH
38104 Magdeburg
BTG Feldberg & Sohn GmbH & Co KG
48365 Bocholt
KG Burspeed Speditionen-GmbH & Co KG
48849 Erfurt
The Hazy

**Zutaten für 4 Portionen**

100 g **Marzipan - Rohmasse**
300 ml Milch
3 Ei(ers)
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)
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
**Freischatzung:** 07.09.06
**Rezept-Statistiken:**
- 12.081 (156)* gelesen
- 148 (0)* gespeichert
- 439 (5)* gedruckt
- 14 (0)* verschickt
- * nur in diesem Monat

**Verfasser:**

**Schlagworte für dieses Rezept**

- Dessert, Mehlspeisen, Süßspeise

**Ähnliche Rezepte**

- Semlor
- Pfaffenhütel-Zitronen
- Spekulatius, gefüllt
- Mandelhappen
- Schoko - Marzipan - Herzen
- Marzipan - Pistazien - Creme
- Pistazienispelfarbe mit Nougatsauce
- Schoko - Marzipan - Eis
- Zwetschgenknödel
- Scheiterhaufen mit Äpfeln und Marzipan

**Rezeptsammlungen**

Dieses Rezept ist in diesen Sammlungen geprüft:

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

R. Schäfer, A. Barbaresi, F. Bildhauer 2013, German Grammar and Linguistics (FU Berlin)
The Hazer
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 \textbf{should not} be in the corpus
  - $1, 2$: document \textbf{should} be in the corpus
  - $0$: undecided/document \textbf{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

<table>
<thead>
<tr>
<th>statistic</th>
<th>early 500</th>
<th>late 500</th>
<th>all 1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>0.566</td>
<td>0.300</td>
<td>0.433</td>
</tr>
<tr>
<td>$\kappa$ (raw)</td>
<td>0.397</td>
<td>0.303</td>
<td>0.367</td>
</tr>
<tr>
<td>$ICC(C, 1)$</td>
<td>0.756</td>
<td>0.679</td>
<td>0.725</td>
</tr>
<tr>
<td>raw ($r \geq 0$)</td>
<td>0.900</td>
<td>0.762</td>
<td>0.831</td>
</tr>
<tr>
<td>raw ($r \geq 1$)</td>
<td>0.820</td>
<td>0.674</td>
<td>0.747</td>
</tr>
<tr>
<td>$\kappa$ ($r \geq 0$)</td>
<td>0.673</td>
<td>0.625</td>
<td>0.660</td>
</tr>
<tr>
<td>$\kappa$ ($r \geq 1$)</td>
<td>0.585</td>
<td>0.555</td>
<td>0.598</td>
</tr>
<tr>
<td>$\kappa$ ($r \geq 2$)</td>
<td>0.546</td>
<td>0.354</td>
<td>0.498</td>
</tr>
</tbody>
</table>
Acceptable results?

- values below 0.68 [Krippendorff, 1980]
- even considering criticism of Krippendorff’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 $\geq 0$:
  A: 78.4%, R: 73.8%, S: 84.9%
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?
Profile development while training

German test profile; $n = 1000$; log10-transformed relative frequencies
Distribution of Badness depending on document length (early profile on early data)

left: DECODW2012; right: UKCOW2012
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
Profile comparison:
Raw agreement between profiles

left: DECOD2012; right: UKCOD2012
DE Coc0212 sample raw profile comparison

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

<table>
<thead>
<tr>
<th></th>
<th>prec</th>
<th>rec</th>
<th>F1</th>
<th>correct</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>0.914</td>
<td>0.959</td>
<td>0.936</td>
<td>0.888</td>
<td>0.849</td>
</tr>
<tr>
<td>A</td>
<td>0.856</td>
<td>0.973</td>
<td>0.911</td>
<td>0.851</td>
<td>0.781</td>
</tr>
<tr>
<td>R</td>
<td>0.808</td>
<td>0.976</td>
<td>0.884</td>
<td>0.811</td>
<td>0.738</td>
</tr>
</tbody>
</table>

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

<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>
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


