

Introduction to Corpus Resources, Annotation and Access: *Web as Corpus*

Sabine Schulte im Walde
Universität des Saarlandes

Heike Zinsmeister
Universität Tübingen

Foundational Course

18th European Summer School in Logic, Language and Information
Málaga, Spain
July 31 - August 4, 2006

Overview

1. Web as Corpus?
2. Build your own corpus from the web

Web as Corpus?



Web as corpus?

The web is immense, free and available by mouse click. It contains hundreds of billions words of text and can be used for all manner of language research.

A corpus is a collection of texts, when viewed as an object of language or literary study.

- 20 terabytes of nonmarkup text searchable by Google in 2003.
- More than 55 million words of Basque indexed by AltaVista in 3/2001.

(Kilgarriff and Grefenstette, 2003)

Google as Query tool and www as corpus?

Objection: Results are not reliable.

- Population and exact hit counts are unknown → no statistics possible.
- Indexing does not allow to draw conclusions on the data.

Google is missing functionalities that linguists / lexicographers would like to have.

Alternative: Use search engine to download data from the net and build a corpus from it.

- known size and exact hit counts → statistics possible.
- people can draw conclusions over the included text types.
- (limited) control over the content.

Web as corpus: Challenges

- Text is embedded in and interspersed with lots of code (**boilerplates**: navigation menus, advertisement, etc.)
- Many languages are present on the web. How to find pages in the wanted language?
- Lack of meta-data
 - author ship
 - language proficiency
 - sublanguage: e-mail, chat, blog
- Legal status
- Character encoding

Build your own corpus from the web

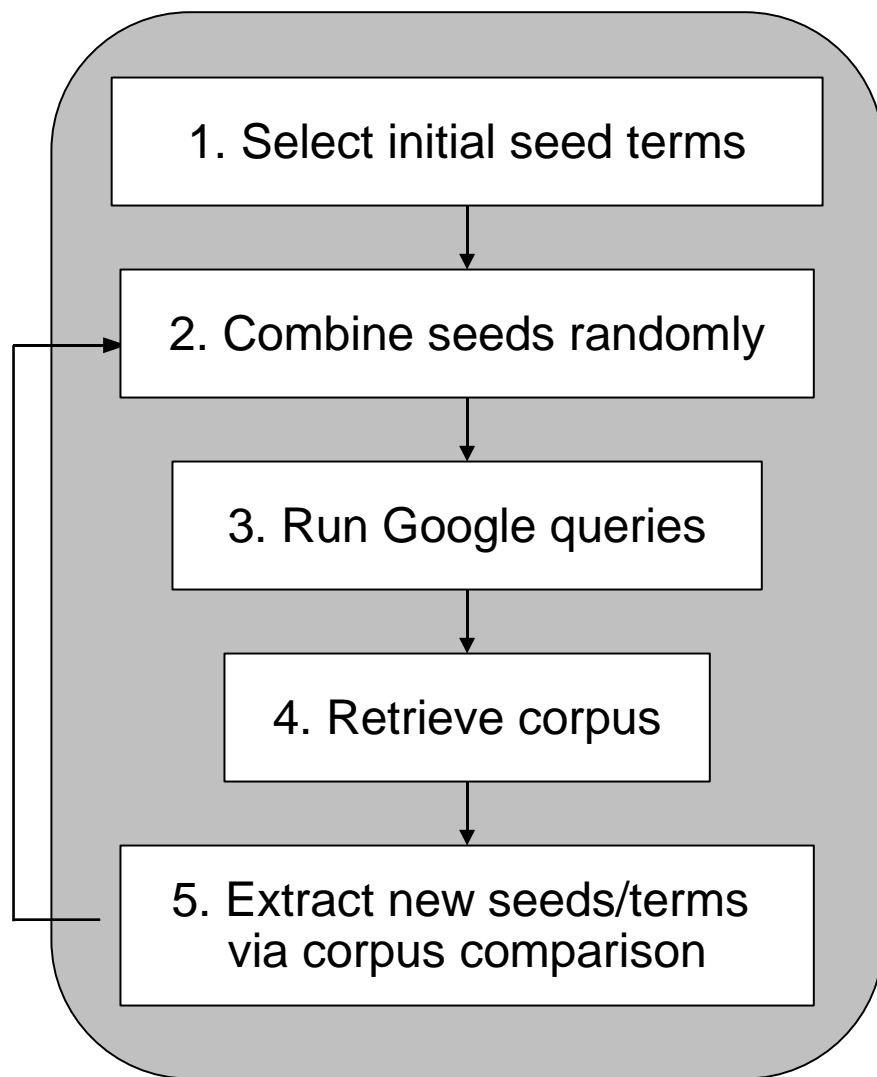
Approaches

- Use Google hit counts.
 - Keller and Lapata, 2003
- Use snippets.
- Use google, then download pages.
 - Baroni and Bernardini, 2004 (BootCat tool)
- Spider from relevant starting sites.
 - Kilgarriff et al., 2006 (Wac toolkit)

BootCat Toolkit

- **Bootstrapping Corpora and Terms from the WWW**
- Uses Google to download pages.
- Iterative knowledge-poor procedure to bootstrap specialized corpora and terms from the WWW.
- Freely available at <http://sslmit.unibo.it/~baroni/>
- A modular set of command-line Perl programs
- Already applied to English, Italian, German, Spanish, Japanese (...)

(slide by Marco Baroni)



Web as Corpus: References

Baroni, Marco and Bernardini, Silvia (2004). “BootCaT: Bootstrapping corpora and terms from the web”. In Proceedings of LREC 2004.

Kilgarriff, Adam and Greffenstette, Gregory (2003). “Introduction to the Special Issue on Web as Corpus”. Tech. Rep. ITRI-03-20, Information Technology Research Institute, University of Brighton. Also published in *Computational Linguistics* 29(3):1–15.

Kilgarriff, Adam, Rundell, Michael, and Dhonnchadha, Elaine U´ı (2005). “Efficient corpus development for lexicography: building the New Corpus for Ireland. In “*Language Resources and Evaluation Journal*”.

Ziai, Ramon and Ott, Niels (2005). “Web as Corpus Toolkit: User’s and Hacker’s Manual”. Lexical Computing Ltd.
<http://www.drni.de/wac-tk/index.php/Documentation>

Web as Corpus: References

- BootCat tools: <http://sslmit.unibo.it/~baroni/bootcat.html>
- BootCat Web tool: <http://:corpora.fi.muni.cz/bootcat/>
- Web as Corpus Workshop at Corpus Linguistics 2005
http://sslmit.unibo.it/~baroni/web_as_corpus_cl05.html
- WaCky project: <http://wacky.sslmit.unibo.it/>

Related tools

- Linguist's Search Engine: <http://lse.umiacs.umd.edu:8080/>
- Kwicfinder: <http://www.kwicfinder.com/KWiCFinder.html>
- Webcorp: <http://www.webcorp.org.uk/>
- WAC toolkit: <http://www.drni.de/wac-tk/>