**COW & texrex: Gigatoken Web Corpora and Tools for Web Corpus Construction**

**Roland Schäfer & Felix Bildhauer**

**Freie Universität Berlin**

http://hsfg.fu-berlin.de/cow/

---

**Goals and Data**

- **Target Corpora**
  - Building on work on the WaCy corpora (Baroni et al., 2009), introducing incremental improvements
  - General purpose and very large, enabling linguistic research of low-frequency phenomena
  - Over 3 billion tokens, better than 10 billion tokens, large enough to drive purpose-specific corpora
  - Best possible random samples from the web (by top-level domain)
  - Mostly free of duplication on the concordance level
  - Containing a considerable amount of quasi-spontaneous and subderivate language (chat, forums, etc.)
  - Languages: German, UK and World English, Castilian, Swedish, French, Italian, Dutch, Danish, Malagasy

- **Software**
  - Full text chain for all core corpus creation including crawler (not including linguistic processing)
  - Independence of search engine results; guaranteed no-cost corpus construction
  - Efficient, cross-platform (written in Objective Pascal with the Fast Pascal compiler), open-source

**Data Collection**

- For custom corpora: long or very long web-crawls using Heritrix 1.4 (similar to Emerson and O’Neil, 2006)
- Seed URLs for Heritrix: search engine results (Yahoo, Bing)
- Maximum number of documents crawled so far for one TLD (DECOW2012): 136,602,410

**Problems with Established Methods**

- BootCat method (Baroni and Bernardini, 2004): relying entirely on search engine results
- WaCy method (Baroni et al., 2009): starting with search engine results and doing short web crawls
- Random sample from the web: impossible; not the same as a random sample from a search engine
- Problem 1: re-use of random samples from a search engine not state-of-the-art in web corpus construction
- Weakness of mid-frequency triple query method (not the methods of Bar-Yossef and Gurevich, 2006)
- Problem 2: shutdown of free search engine APIs for massive URL requests
- Problem 3: host bias: search engines and short (breadth-first) crawls leading to samples unecessarily biased towards certain web hosts

**Connected Text Recognition**

- Problem not simply foreign language documents, but tag clouds, lists, tables, etc.
- Simple language identification (a textbook matter) insufficient
- WaCy method: requiring certain type and token counts of function words
- Problem depends heavily on document length, is unreliable for long documents with mixed text
- Our solution: calculate the standardized sum of negative deviation \( R(D) \) of frequencies of certain function words per document \( d \) compared to an \( f \) of training documents
- \( \alpha(x) \): a weighted mean for \( x \) in \( D \), \( \beta(x) \): corresponding weighted standard deviation
- For unknown document \( d \) in production form, \( f(D) \): relative frequency of \( x \)

\[
\alpha(x) = \frac{\sum_{x \in D} \rho(x)}{|D|}
\]

\[
\beta(x) = \sqrt{ \sum_{x \in D} (\rho(x) - \alpha(x))^2 / (|D| - 1) } \]

\[
R(D) = \frac{\sum_{x \in D} (\rho(x) - \alpha(x))}{\beta(x)}
\]

- For COW corpus: removal of documents with \( R(D) > 10 \) (very strict, so: Recall < 0.95, Precision > 0.95)

**Evaluation**

- **Duplication**
  - Reduction in number of documents by w-shingling (5% or higher w-shingling overlap): notice remaining duplication in WaCy corpora due to “simplification” of shingling algorithm.
  - Assessment of remaining duplication:
    - Choose frequent word \( W \)
    - Examine all occurrences of \( W \) plus \( \sigma \) characters to its left and right: count repetitions
    - No distinction between citations and duplicates: multiple counts per document possible; yet conservative estimate of duplication

- **Quality of Connected Text Detection**
  - As a simple language identifier on 105 German and 13 non-German/dialect test documents:
    - Precision = 1, Recall = 0.97, F = 0.99
    - Promising results as connected text identifier, but problem of defining what a good document is
    - Goal: provide a set of training and text documents to define a gold standard
    - Right \( R(D) \) for synthetic German test documents of different sizes containing different amounts of tag cloud material

**Genres and Text Types**

- Classification scheme based on Shamir (2006), with modifications
- Manual coding of 200 documents per corpus
- Substantially to almost perfect inter-coder agreement (measured for German corpus only)
- CEAS for 90% confidence level, \( n = 200 \)

**Goals and Data**

- **Target Corpora**
  - Building on work on the WaCy corpora (Baroni et al., 2009), introducing incremental improvements
  - General purpose and very large, enabling linguistic research of low-frequency phenomena
  - Over 3 billion tokens, better than 10 billion tokens, large enough to drive purpose-specific corpora
  - Best possible random samples from the web (by top-level domain)
  - Mostly free of duplication on the concordance level
  - Containing a considerable amount of quasi-spontaneous and subderivate language (chat, forums, etc.)
  - Languages: German, UK and World English, Castilian, Swedish, French, Italian, Dutch, Danish, Malagasy

- **Software**
  - Full text chain for all core corpus creation including crawler (not including linguistic processing)
  - Independence of search engine results; guaranteed no-cost corpus construction
  - Efficient, cross-platform (written in Objective Pascal with the Fast Pascal compiler), open-source

**Data Collection**

- For custom corpora: long or very long web-crawls using Heritrix 1.4 (similar to Emerson and O’Neil, 2006)
- Seed URLs for Heritrix: search engine results (Yahoo, Bing)
- Maximum number of documents crawled so far for one TLD (DECOW2012): 136,602,410

**Problems with Established Methods**

- BootCat method (Baroni and Bernardini, 2004): relying entirely on search engine results
- WaCy method (Baroni et al., 2009): starting with search engine results and doing short web crawls
- Random sample from the web: impossible; not the same as a random sample from a search engine
- Problem 1: re-use of random samples from a search engine not state-of-the-art in web corpus construction
- Weakness of mid-frequency triple query method (not the methods of Bar-Yossef and Gurevich, 2006)
- Problem 2: shutdown of free search engine APIs for massive URL requests
- Problem 3: host bias: search engines and short (breadth-first) crawls leading to samples unecessarily biased towards certain web hosts

**Connected Text Recognition**

- Problem not simply foreign language documents, but tag clouds, lists, tables, etc.
- Simple language identification (a textbook matter) insufficient
- WaCy method: requiring certain type and token counts of function words
- Problem depends heavily on document length, is unreliable for long documents with mixed text
- Our solution: calculate the standardized sum of negative deviation \( R(D) \) of frequencies of certain function words per document \( d \) compared to an \( f \) of training documents
- \( \alpha(x) \): a weighted mean for \( x \) in \( D \), \( \beta(x) \): corresponding weighted standard deviation
- For unknown document \( d \) in production form, \( f(D) \): relative frequency of \( x \)

\[
\alpha(x) = \frac{\sum_{x \in D} \rho(x)}{|D|}
\]

\[
\beta(x) = \sqrt{ \sum_{x \in D} (\rho(x) - \alpha(x))^2 / (|D| - 1) } \]

\[
R(D) = \frac{\sum_{x \in D} (\rho(x) - \alpha(x))}{\beta(x)}
\]

- For COW corpus: removal of documents with \( R(D) > 10 \) (very strict, so: Recall < 0.95, Precision > 0.95)

**Evaluation**

- **Duplication**
  - Reduction in number of documents by w-shingling (5% or higher w-shingling overlap): notice remaining duplication in WaCy corpora due to “simplification” of shingling algorithm.
  - Assessment of remaining duplication:
    - Choose frequent word \( W \)
    - Examine all occurrences of \( W \) plus \( \sigma \) characters to its left and right: count repetitions
    - No distinction between citations and duplicates: multiple counts per document possible; yet conservative estimate of duplication

- **Quality of Connected Text Detection**
  - As a simple language identifier on 105 German and 13 non-German/dialect test documents:
    - Precision = 1, Recall = 0.97, F = 0.99
    - Promising results as connected text identifier, but problem of defining what a good document is
    - Goal: provide a set of training and text documents to define a gold standard
    - Right \( R(D) \) for synthetic German test documents of different sizes containing different amounts of tag cloud material

**Genres and Text Types**

- Classification scheme based on Shamir (2006), with modifications
- Manual coding of 200 documents per corpus
- Substantially to almost perfect inter-coder agreement (measured for German corpus only)
- CEAS for 90% confidence level, \( n = 200 \)