The academic Web-as-Corpus

*Introducing the acWaC-EU corpus*

Adriano Ferraresi
Silvia Bernardini

Web as Corpus Workshop@CL2013
UCREL, Lancaster University
22 July 2013
Outline

• Background
  – object of study
  – previous work

• Corpus
  – construction
  – evaluation

• Case studies
  – (semi-)modal verbs (usage-oriented)
  – naïve text classification (methodology-oriented)

• Plans for the future
Object of study

• Institutional academic English…
  – texts used by higher education institutions for everyday communication
    • e.g. mission statements, news, course catalogues
  – as opposed to disciplinary genres
    • e.g. research articles, book reviews, grant proposals

• … beyond the native(-only) standard
  – “In order to understand the use of English in present-day academic communities, it is vital to look at English as a lingua franca” (Mauranen 2010:6-7)
  – academic modules/degree courses in English are essential for internationalization (Altbach and Knight 2007)
    ➢ Bologna Translation service (Depraetere et al 2011)
Previous work

• Critical Discourse Analysis
  – Marketization of university discourse (Fairclough 1993, Swales 2004)
  – Universities “have adopted the language of business and industry, managerialism and neoliberalism” (Morrish and Sauntson 2013:78)

• Corpus linguistics
  – TOEFL 2000 – Spoken and Written Academic Language Corpus (T2K-SWAL)
  – Michigan Corpus of Academic Spoken English (MICASE)

• Web-as-Corpus linguistics
  – Crawls of academic (native) English websites (Thelwall 2005, Rehm 2002)
    • mainly for genre classification/web document clustering
  – Automatic construction of parallel corpora (Resnik and Smith 2003)
Why acWaC-EU?

- **Descriptively**, to compare native and ELF textual practices across EU countries
- **Methodologically**, to establish practices for building WaC ELF corpora
- **Practically**, to provide resources for writers/translators (in native and ELF countries)
Building acWaC-EU (ELF) or finding a few needles in a huge haystack...

- List of EU Universities from http://www.webometrics.info
- Look for English-language homepage (if any)
  - `<a>` tags with `english|eng|en` in `href`, `class`, `title` and in link text
  - Precision: 84%
  - HTML header: `lang/content` attributes set to `en`, `en-US` or `en-GB`
  - Manual check of these pages (precision: 26.3%)
Building acWaC-EU (ELF) or finding a few needles in a huge haystack…

- Seed URL retrieval
- Harvesting of pages
- Cleaning, annotation and indexing

- Download all pages linked from (English) homepage
- two levels of recursion
- HTML only
Building acWaC-EU (ELF) or finding a few needles in a huge haystack…

- Language identification, boilerplate stripping and de-duping algorithms developed for WaCky corpora
- Part-of-speech tagging / lemmatization with TreeTagger
- Indexing with Corpus WorkBench
- Meta-data encoded
  - URL from which text was downloaded
  - level of recursion (0 to 2)
  - ELF/NAT status
  - University name / country / EU27-non EU27 / rank
  - L1 family (Germanic, Romance, Slavic, …)
Corpus stats

<table>
<thead>
<tr>
<th></th>
<th>ELF</th>
<th>Native</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>41 mln</td>
<td>46 mln</td>
<td>87 mln</td>
</tr>
<tr>
<td>Texts</td>
<td>73 K</td>
<td>68 K</td>
<td>141 K</td>
</tr>
<tr>
<td>Unis</td>
<td>2,2 K</td>
<td>~ 300</td>
<td>2,5 K</td>
</tr>
<tr>
<td>Countries</td>
<td>46</td>
<td>4</td>
<td>50</td>
</tr>
</tbody>
</table>

Number of tokens (%) by main language families

- romance 13%
- slavic 13%
- germanic-ELF 12%
- uralic 4%
- turkic 2%
- baltic 1%
- hellenic 1%
- semitic 1%
- germanc-NAT 55%
Evaluating the method

- Experiment
  - acWaC-EU vs. Baseline method
    - Identify EN home and download pages linked from there vs. download all pages linked from home (in national language)
    - 3 levels of recursion
  - 33 Uni’s from 3 ELF countries
    - Serbia, Spain and Sweden
<table>
<thead>
<tr>
<th>Method</th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>acWaC method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downl.</td>
<td>73</td>
<td>3,771</td>
<td>42,070</td>
<td>275,638</td>
</tr>
<tr>
<td>Final</td>
<td>22</td>
<td>937</td>
<td>5,818</td>
<td>12,318</td>
</tr>
<tr>
<td>RATIO</td>
<td>30.1%</td>
<td>24.8%</td>
<td>13.8%</td>
<td>4.4%</td>
</tr>
<tr>
<td><strong>Baseline method</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downl.</td>
<td>99</td>
<td>6,470</td>
<td>70,605</td>
<td>486,900</td>
</tr>
<tr>
<td>Final</td>
<td>0</td>
<td>133</td>
<td>2,396</td>
<td>12,767</td>
</tr>
<tr>
<td>RATIO</td>
<td>0.0%</td>
<td>2.1%</td>
<td>3.4%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>
Corpus evaluation

• Sample of
  – 99 pages: Nat
  – 99 pages: ELF
    • 33 Germanic (ELF), 33 Romance, 33 Slavic
  – Categorization in terms of broad topics/genres
Case study 1 - using the corpus

- Modal and semi-modal verbs
  - “by far the most common grammatical device used to mark stance in university registers” (Biber 2006:95)
  - in Nat vs. ELF texts
**Frequency of modals in the NAT vs. ELF sub-corpus**

<table>
<thead>
<tr>
<th>(Semi-) modal</th>
<th>p (Fisher)</th>
</tr>
</thead>
<tbody>
<tr>
<td>can</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>could</td>
<td>ns</td>
</tr>
<tr>
<td>have to</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>may</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>might</td>
<td>ns</td>
</tr>
<tr>
<td>must</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>need to</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>shall</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>should</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>will</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>would</td>
<td>ns</td>
</tr>
</tbody>
</table>
Frequency of modals in the NAT vs. ELF sub-corpus

<table>
<thead>
<tr>
<th>(Semi-)modal</th>
<th>p (Fisher)</th>
</tr>
</thead>
<tbody>
<tr>
<td>can</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>could</td>
<td>ns</td>
</tr>
<tr>
<td>have to</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>may</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>might</td>
<td>ns</td>
</tr>
<tr>
<td>must</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>need to</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>shall</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>should</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>will</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>would</td>
<td>ns</td>
</tr>
</tbody>
</table>
Case study 1 - using the corpus

- Modals and semi-modal verbs
  - “by far the most common grammatical device used to mark stance in university registers” (Biber 2006:95)
  - in Nat vs. ELF texts
  - In different language families
Case study 1 - using the corpus

- Modals and semi-modal verbs
  - “by far the most common grammatical device used to mark stance in university registers” (Biber 2006:95)
  - in Nat vs. ELF texts
  - In different language families
  - *Shall*: a qualitative perspective
    - *(PP | NOUN) + shall + VB*
    - Nat: formal/regulatory, e.g. “personal data shall be processed”
      - (also: first person expression of volition, e.g. “we shall be offering”)
    - ELF-Romance: like Native, e.g. “litigation shall come”
    - ELF-Germanic and ELF-Slavic: formal but not regulatory, e.g. “supervisor shall be employed”
Case study 2 – future perspectives

• Naïve text classification for subcorpus construction based on URLs
  – Frequency list for slash-separated parts of URLs without transfer protocols and domain names. E.g.
    - http://apps.uc.pt/courses/EN/course/1514
Case study 2 – future perspectives

- Naïve text classification for subcorpus construction based on URLs

  - Frequency list for slash-separated parts of URLs without transfer protocols and domain names. E.g.
    - http://apps.uc.pt/courses/EN/course/1514

  - 8488 news
  - 5903 courses
  - 5170 research
  - 3976 english
  - 3331 pages
  - 2759 about
  - 2508 study
  - 2139 undergraduate
  - 2098 2013
  - 2055 content
Case study 2 – future perspectives

- Naïve text classification for subcorpus construction based on URLs
  - Frequency list for slash-separated parts of URLs without transfer protocols and domain names. E.g.
    - http://apps.uc.pt/courses/EN/course/1514

<table>
<thead>
<tr>
<th></th>
<th>news</th>
<th>courses</th>
<th>research</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELF</td>
<td>3,673,205</td>
<td>800,487</td>
<td>1,901,098</td>
</tr>
<tr>
<td>NAT</td>
<td>5,488,887</td>
<td>7,576,082</td>
<td>2,566,095</td>
</tr>
</tbody>
</table>

(Number of tokens by subcorpus)
Case study 2 – future perspectives

- 50 pages per keyword per subcorpus (ELF vs. NAT)
- Courses
  - 90% describe courses
  - 10% regulations or facilities of courses
- News
  - 100% news about academic events, partnerships, discoveries
- Research
  - 99% groups, findings, projects, grants, infrastructure, support, staff profiles, homepages of institutes
Plans for the future

• Test efficacy of method for building topic/genre-restricted subcorpora based on URL syntax
• Make the corpus available as a set of N-grams
• Go global: extend the crawl to university websites from other continents

• … for details on acWaC-EU and future updates: http://mrscoulter.sslmit.unibo.it/acwac
The academic Web-as-Corpus

*Introducing the acWaC-EU corpus*

THANKS

adriano.ferraresi@unibo.it
silvia.bernardini@unibo.it

Web as Corpus Workshop@CL2013
UCREL, Lancaster University
22 July 2013