ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

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

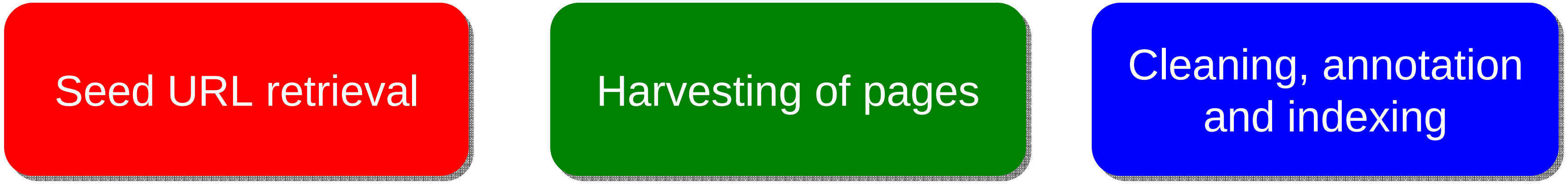
- Critical Discourse Analysis
  - Marketization of university discourse (Fairclough 1993, Swales 2004)
  - Universities “have adopted the language of business and industry, managerialism and neoliberalism” (Morrish and Sauntson 2013:78)
- Corpus linguistics
  - TOEFL 2000 – Spoken and Written Academic 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...



Seed URL retrieval

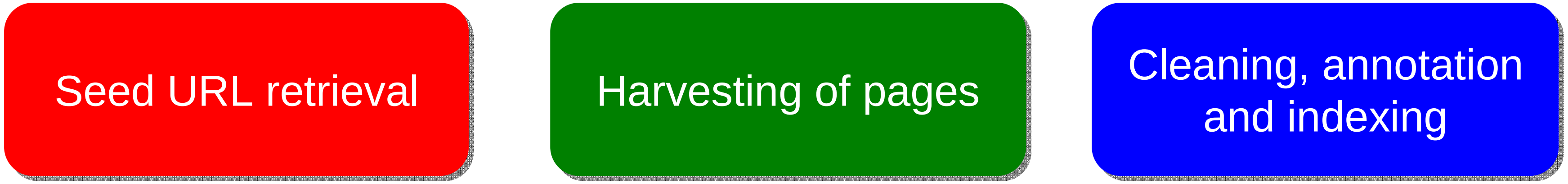
Harvesting of pages

Cleaning, annotation and indexing

- 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...



- 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…

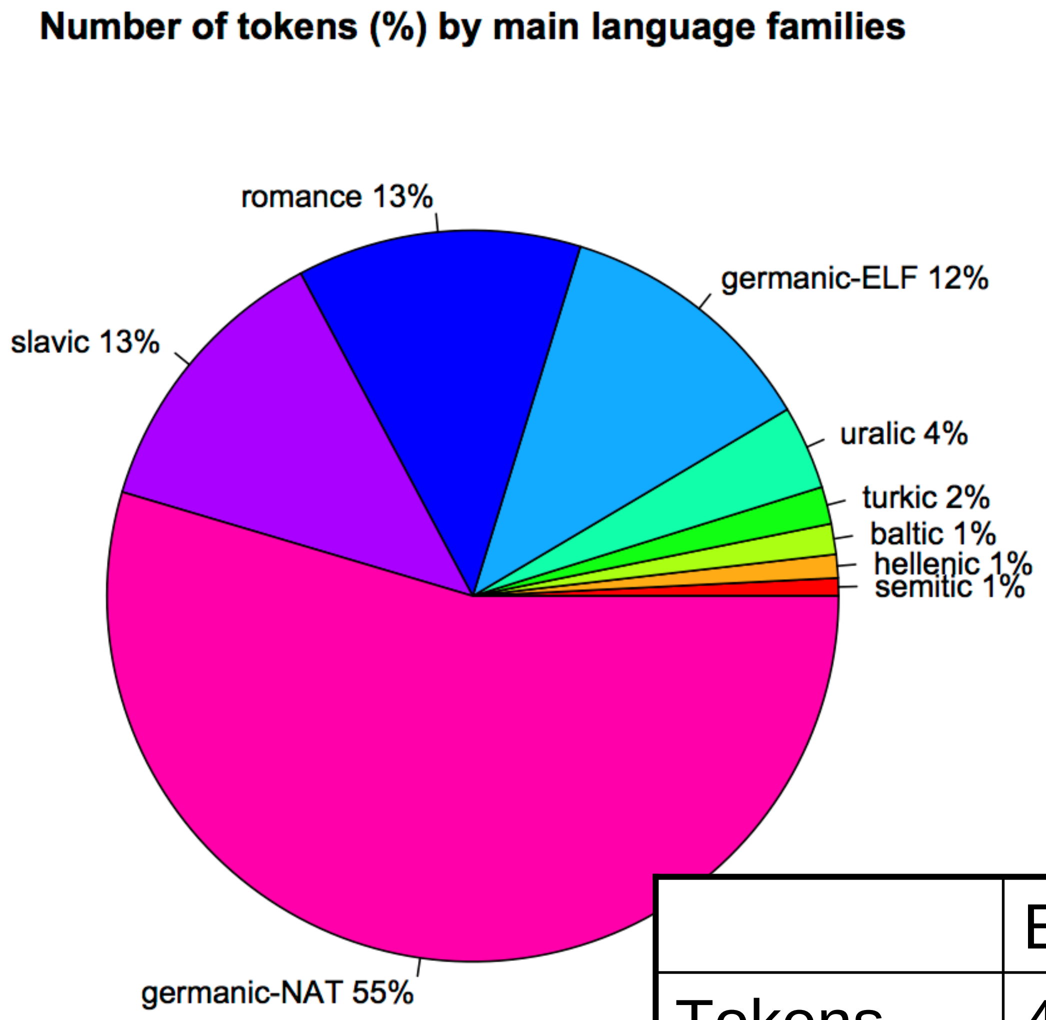
Seed URL retrieval

Harvesting of pages

Cleaning, annotation and indexing

- Language identification, boilerplate stripping and de-dupe 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



	ELF	Native	TOTAL
Tokens	41 mln	46 mln	87 mln
Texts	73 K	68 K	141 K
Unis	2,2 K	~ 300	2,5 K
Countries	46	4	50

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

		Level 0	Level 1	Level 2	Level 3
acWaC method	Downl.	73	3,771	42,070	275,638
	Final	22	937	5,818	12,318
	RATIO	30.1%	24.8%	13.8%	4.4%
Baseline method	Downl.	99	6,470	70,605	486,900
	Final	0	133	2,396	12,767
	RATIO	0.0%	2.1%	3.4%	2.6%

Corpus evaluation

- Sample of
  - 99 pages: Nat
  - 99 pages: ELF
    - 33 Germanic (ELF), 33 Romance, 33 Slavic
  - Categorization in terms of broad topics/genres

Course descriptions

Facilities

General / welcoming texts

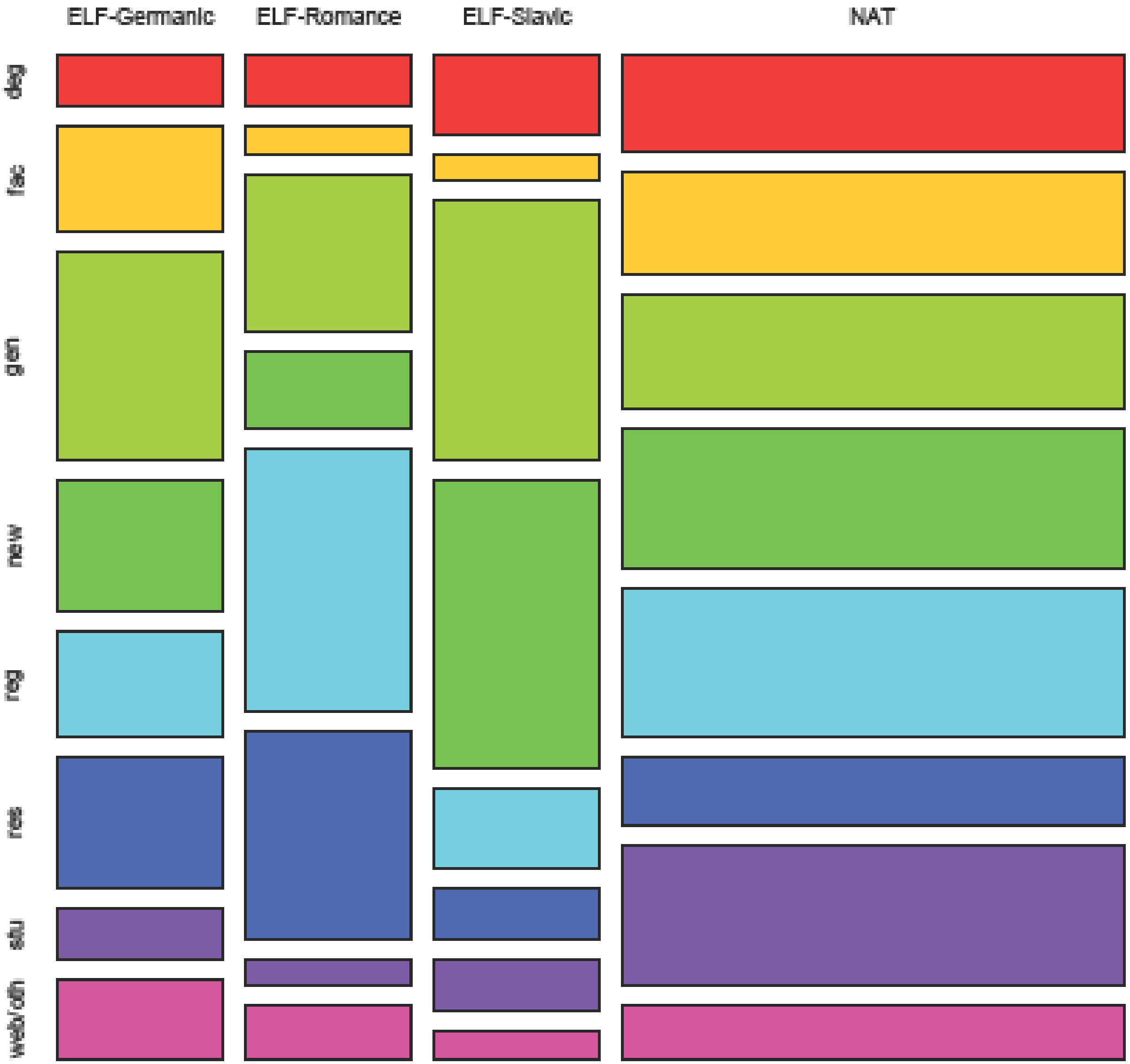
News and announcements

Regulations

Research-related

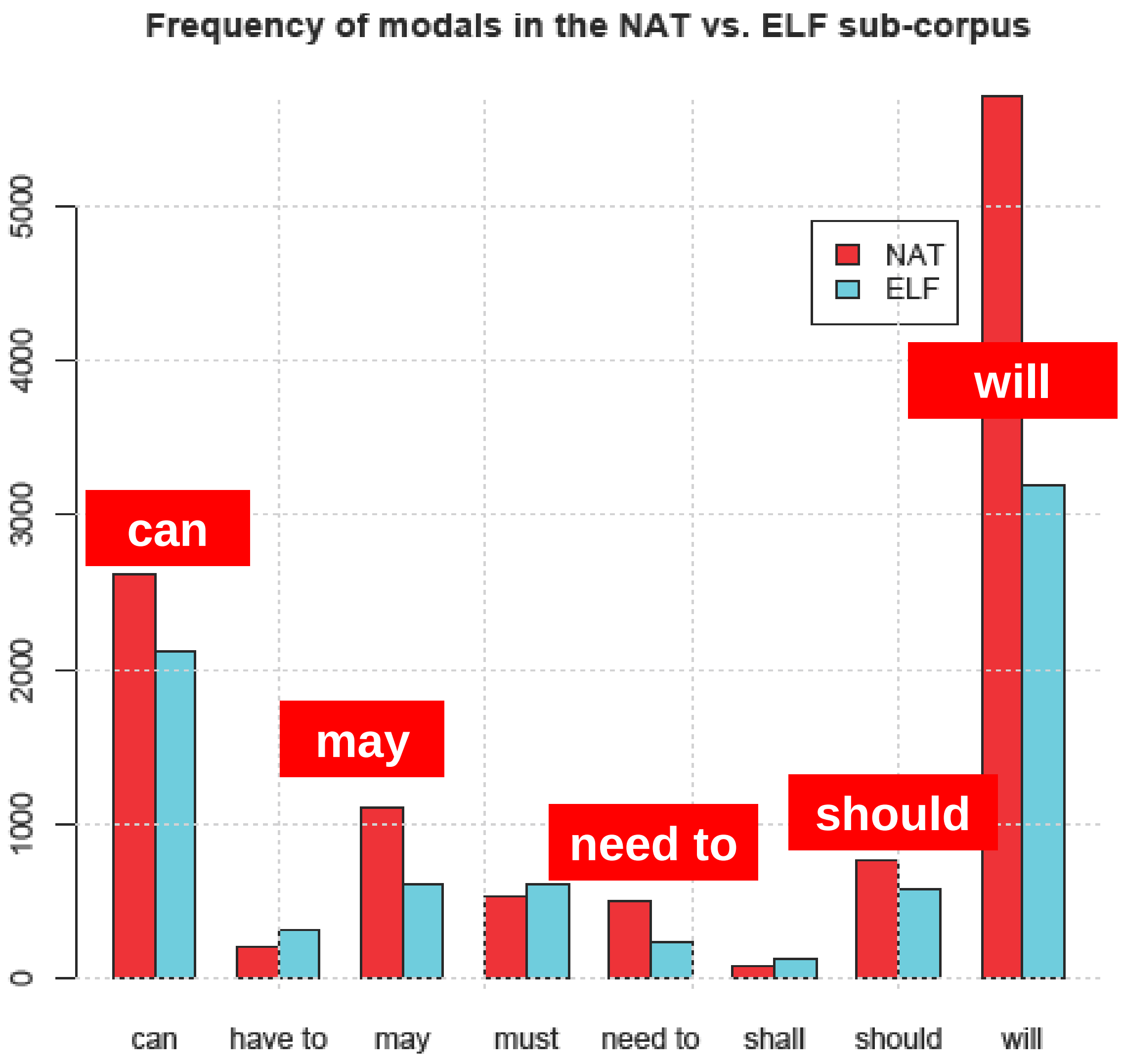
Student life

Other



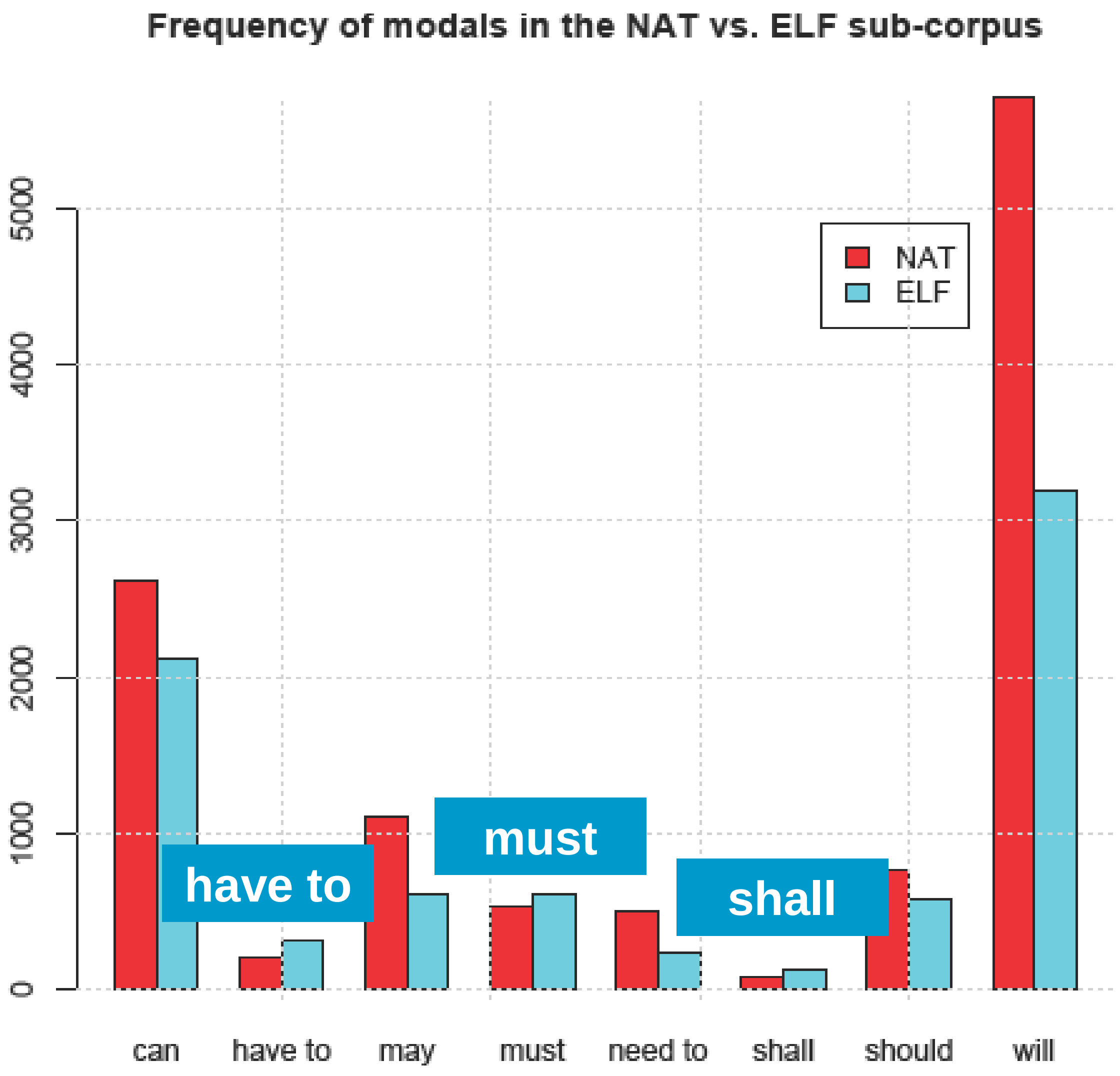
Case study 1 - using the corpus

- 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



(signifcant only)

(Semi-) modal	p (Fisher)
can	< 0.001
could	*ns*
have to	< 0.001
may	< 0.001
might	*ns*
must	< 0.001
need to	< 0.001
shall	< 0.001
should	< 0.05
will	< 0.001
would	*ns*

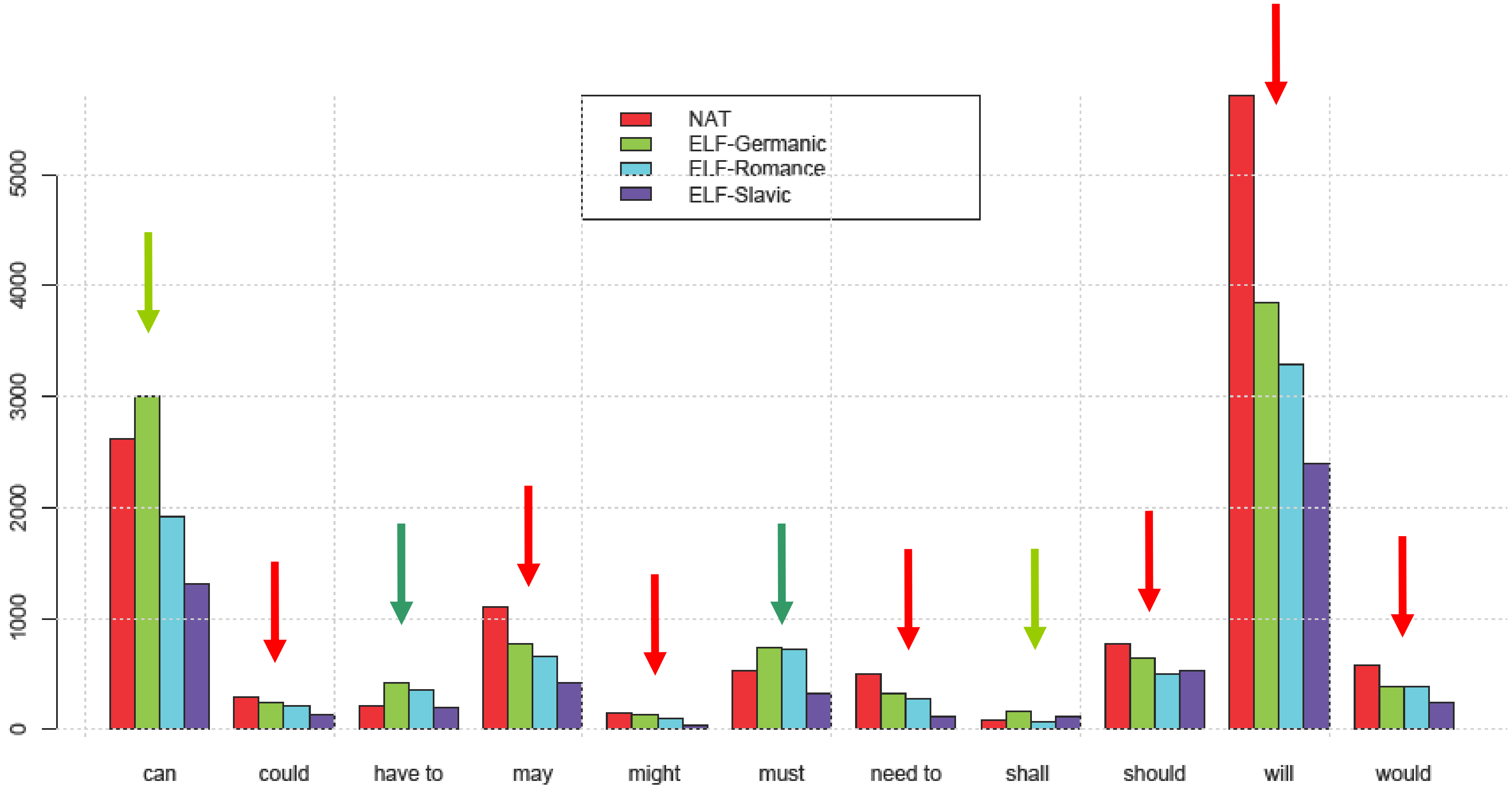


(signifcant only)

(Semi-) modal	**p (Fisher)**
can	< 0.001
could	*ns*
have to	< 0.001
may	< 0.001
might	*ns*
must	< 0.001
need to	< 0.001
shall	< 0.001
should	< 0.05
will	< 0.001
would	*ns*

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



can could have to may might must need to shall should will would

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://www.essex.ac.uk~~/news/event.aspx?e_id=5059
  - ~~http://recherche.isae.fr~~/en/research/scientific_policy/issues.html
  - ~~http://apps.uc.pt~~/courses/EN/course/1514

Case study 2 – future perspectives

- Naïve text classification for subcorpus construction based on URLs
  - Frequenc arts of URLs without tr n names. E.g.
  - ~~http://www.c~~ 9
  - ~~http://reche~~ _policy/issues.html
  - ~~http://apps.~~

```
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://www.essex.ac.uk~~/news/event.aspx?e_id=5059
  - ~~http://recherche.isae.fr~~/en/research/scientific_policy/issues.html
  - ~~http://apps.uc.pt~~/courses/EN/course/1514

	news	*courses*	*research*
ELF	3,673,205	800,487	1,901,098
NAT	5,488,887	7,576,082	2,566,095

(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

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

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