Genre classification for a corpus of academic webpages

Erika Dalan, Serge Sharoff

University of Bologna, University of Leeds

12 August 2016

UNIVERSITY OF LEEDS

Institutional-academic English

- Language of international communication

BUT British and Irish universities vs
ELF (English as Lingua Franca)

- Any difference in the language choices made?
  e.g. Modals of permission, possibility and ability
  (Bernardini, Ferraresi, 2013)
- Any influence from the source language?
- University websites do not correspond to **a** genre:
  descriptions, narratives, instructions, information and opinions

UNIVERSITY OF LEEDS

Corpus collection

- Ten from top 30% of universities in each European country from QS World University Rankings
  `http://www.topuniversities.com/qs-world-university-rankings`
- Corpus building consists of three steps:
  1. retrieving a list of seed URLs (English homepages);
  2. crawling university websites starting from the list of URLs;
  3. post-processing data, annotation and indexing.

	ELF	Native EN	Total
Tokens	9,375,739	11,813,692	21,189,431
Texts	17,383	17,562	34,945
Universities	78	13	91
Countries	27	2	29

UNIVERSITY OF LEEDS

Functional Text Dimensions (FTDs)

A7, instruct	To what extent does the text aim at teaching the reader how something works? (FAQ)
A8, hardnews	To what extent does the text appear to be an informative report of recent events? (Newsitem)
A9, legal	To what extent does the text lay down a contract or specify a set of regulations? (T&C)
A12, promo	To what extent does the text promote a product or service?
A14, academ	To what extent does the text serve as an example of academic research?
A16, info	To what extent does the text provide information to define a topic? (Encyclopedic article)
A21, narrate	To what extent does the text describe a chronologically ordered sequence of events? (Story)

UNIVERSITY OF LEEDS

Topology in action

About us webpages

some webpages are strictly informational (**A16**)
some are narrative (**A21**)
some combine information (**A16**) with promotion (**A12**)

- 897 webpages, randomly sampled from the main corpus for training
- Character n-grams for predicting amounts of each FTD
- Models applied to the rest of the corpus

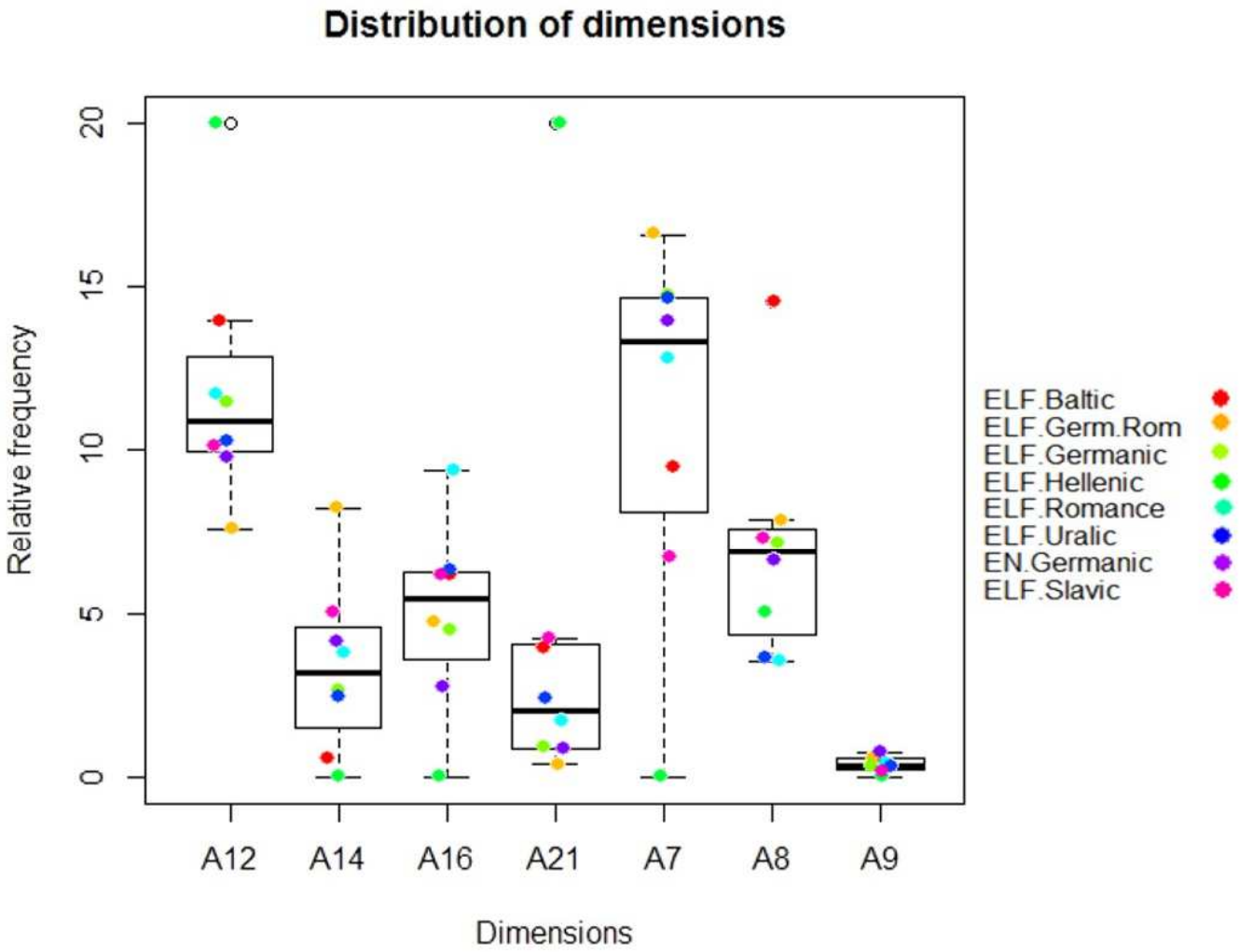
UNIVERSITY OF LEEDS

Classification results

	A7	**A8**	**A9**	**A12**	**A14**	**A16**	**A21**
% training	8.4	5.0	3.2	8.5	6.3	13.6	5.5
F-measure (CV)	0.95	0.92	0.96	0.85	0.93	0.79	0.94
% total corpus	13.9	6.3	0.5	10.2	3.9	3.3	1.1
N. of pages	4,737	2,168	190	3,492	1,353	1,127	383

UNIVERSITY OF LEEDS

Analysis of language varieties



UNIVERSITY OF LEEDS

Conclusions

- Feasible to collect a large corpus and classify it for genres automatically with reasonable accuracy using character n-grams
- Instructional and promotional webpages are most common
- Variation between the countries:
  Greece more pages on university history (narrative)
  Switzerland more pages on research (academic)
- More linguistic research needed comparing linguistic constructions in native English and ELF for each FTD
  Evaluative expressions in info-promotional pages;
  delivery of instructions, etc
- Language teaching for students and education for ELF writers
- Relationship between university rankings and text types

UNIVERSITY OF LEEDS