Genre classification for a corpus of academic webpages

Erika Dalan
University of Bologna
erika.dalan@unibo.it

Serge Sharoff
University of Leeds
s.sharoff@leeds.ac.uk

Abstract

In this paper we report our analysis of the similarities between webpages that are crawled from European academic websites, and comparison of their distribution in terms of the English language variety (native English vs English as a lingua franca) and their language family (based on the country's official language). After building a corpus of university webpages, we selected a set of relevant descriptors that can represent their text types using the framework of the Functional Text Dimensions. Manual annotation of a random sample of academic pages provides the basis for classifying the remaining texts on each dimension. Reliable thresholds are then determined in order to evaluate precision and assess the distribution of text types by each dimension, with the ultimate goal of analysing language features over English varieties and language families.

1 Introduction

English is increasingly regarded as the language of international communication in professional and institutional settings. In particular, it is the main language used by the European universities to communicate to their audience outside of their own country. English language communication is both a strategic choice for enhancing competitiveness and prestige, with the ultimate goal of attracting international students, and a transparency requirement imposed by the European Higher Education Area (EHEA).[1] At the same time, one can expect that the strategies used for communication vary according to culture and language factors. For instance, British and Irish universities may adopt specific practices that differ from the ones of their counterparts on the continent, which are likely to be using ELF, English as Lingua Franca (Mollin, 2006). Differences may occur on at least two levels. First, on the higher level of genres and second, on the level of language patterns that are used to fulfil specific communicative functions. As regards the former, and with reference to university websites, related work has mainly focused on single genres, rather than the whole website. Some of these genres include *About us* pages (Caiazzo, 2011), Academic Course Descriptions or ACDs (Gesuato, 2011), international student prospectuses (Askehave, 2007), module descriptions (Bernardini et al., 2010) and mission statements (Morrish and Sauntson, 2013). Fewer studies have described university websites as a standalone unit, probably because of their high variability in terms of text types and genres. Based on a case study carried out on a small sample of universities (Dalan, 2015), both native English and ELF websites comprise five main textual functions - i.e. desctiptions, narratives, instructions, information and opinions – and a set of more structured genres such as FAQs, news and news archives, forums, descriptions of research projects, personal homepages (PHPs) and many others. Furthermore, some texts belong to proper academic domains (e.g. research papers and abstracts), others to institutional domains (the vast majority of running text) and others are derived from professional settings following the marketization of higher education (e.g. testimonials and *Why choose us* pages).

This wealth of genres and text types makes university websites a sort of a colony of genres that deserves to be further studied in terms of its textual functions.

As for language choices, Saichaie (2011) has investigated university websites using critical dis-

[1] http://www.ehea.info/.

Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task, pages 90–98,
Berlin, Germany, August 7-12, 2016.

course analysis. By analysing a sample of 12 US colleges, he notes a standardisation in the use of promotional language practices, in such a way that generic images tend to be delivered, regardless of how prestigious universities are. Ferraresi and Bernardini (2013) conducted a case study on the use of modal and semi-modal verbs by academic institutions in Europe and noted that native English texts show higher frequencies of modal verbs as compared to ELF university webpages. Modals of permission, possibility and ability seem to be used more widely in native texts as compared to ELF texts. It is still unclear whether these observations may be related to other variables as well, such as the set of genres mentioned above. Different institutional practices between native English and ELF countries may influence the quality and quantity of pages associated with specific functions. Therefore, finding a reliable method for classifying academic pages may help overcome or minimize biases related to genre variability. Automatic classification of university web-based genres is a fundamental preliminary step for comparing native English and ELF language patterns, as well as a thriving research area in itself that needs to be further explored.

In this paper, we will discuss the methods used for corpus collection (Section 2), a typology used for classifying our texts (Section 3), present the experimental setup (Section 4), analyse the results (Section 5) and discuss further research directions (Section 6).

2 Corpus collection

As mentioned in the Section 1, the final aim of this corpus is to compare communicative strategies of ELF and native English countries in university websites. Due to a lack of standards and best-practices as regards translation, localisation or drafting of online contents in English within the ELF community (Costales, 2012; Palumbo, 2013), only high-ranked universities are considered for inclusion in the corpus, in the attempt of obtaining a golden sample. Furthermore, texts in the gold standard are more easily comparable considering that these universities are evidently involved in the international scene. Therefore, a few design criteria were defined to collect a sample of academic webpages. Criteria for corpus building include the full list of European countries and a selection of universities based on the total number of universities per country listed in the QS World University Rankings.[2] The top 30% of universities in each country was chosen, fixing a maximum of ten. The procedure for text collection followed the pipeline described in the acWaC project (Bernardini and Ferraresi, 2013), including post-processing techniques developed in the WaCky project (Baroni and Bernardini, 2006). Corpus building consists of three steps: a) retrieving a list of seed URLs, i.e. university English homepages; b) crawling university websites starting from the list of URLs; c) post-processing data, annotation and indexing.

As concerns the first step, due to the relatively limited number of universities included in this corpus, English homepages of ELF universities were identified manually. The list of URLs was then used to run a crawl of university websites, starting from homepages down to level two, by following webpages internal links. The third step includes boilerplate removal, de-duplication and language identification. The whole process discarded 10% of universities overall, either because homepages could not been fetched or because they were removed during language identification processes. A set of metadata was also defined, in order to account for internal categorisation and to register contextual information. The list of metadata comprises:

- webpage URL and university English homepage;
- university extended name and main domain;
- QS World University overall ranking and QS World University score associated with the number of international students;
- status (public/private) and size (s/m/l/xl), as registered in the ranking;
- family of the country official language (e.g. Germanic in Norway and Romance in Italy);
- variety of English (either native in the UK and Ireland or ELF);
- level of crawling (from 0 to 2, where 0 is the homepage).

The final corpus contains approximately 20M tokens and 35K texts produced in 91 universities, 78 of which represent ELF countries whereas 13 represent the countries with native English. Table 1 and Table 2 provide descriptive statistics of the final corpus, split by language variety and

[2] http://www.topuniversities.com/qs-world-university-rankings

	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

Table 1: Corpus statistics by English language variety (ELF and native English).

language family (Table 2 refers to ELF countries only).

3 Text typology

The webpages in the corpus can express several functions at the same time. For example, typical *About us* pages include informative descriptions, 'Description of a thing' according to the Web text classification scheme (Egbert et al., 2015), as well as promotional materials ('Informational Persuasion'). In order to deal with such variation we adapted the typology based on Functional Text Dimensions (FTD) (Forsyth and Sharoff, 2014) by selecting the following dimensions relevant to the academic webpages collected for this study:

A7, instruct To what extent does the text aim at teaching the reader how something works?

A8, hardnews To what extent does the text appear to be an informative report of events recent at the time of writing?

A9, legal To what extent does the text lay down a contract or specify a set of regulations?

A12, compuff To what extent does the text promote a product or service?

A14, scitech 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?

A21, narrate To what extent does the text describe a chronologically ordered sequence of events?

Application of this procedure leads to a compact description of each text as scoring on some of the dimensions. For example, some *About us* webpages are strictly informational (**A16**),[3] some are narrative (**A21**),[4] while others combine information with promotion.[5]

We have annotated a subset of 897 webpages, randomly sampled from the main corpus. Due to limited resources, annotation was done by one annotator only. However, other studies which used the FTD annotation categories listed above demonstrated reasonable interannotator agreement levels, with Krippendorff's α ranging from 0.78 to 0.97 for different FTDs (Sharoff, 2015).

Sampling was done by selecting the ten pages for each university randomly.[6] To balance the lack of information required to perform a stratified sample and the need for a representative sample of most text types, we have manually analysed URLs to make sure that specific portions of the website did not dominate over other portions. If URLs were skewed towards a portion of a website (e.g. www.bg.ac.rs/en/bodies/), more pages were taken from other uncovered sections. Each webpage was annotated using a scale from 0 to 2, with 0 meaning that the descriptor is not present at all, 0.5 meaning that it is present to a small extent, 1 meaning that it is partly present and 2 meaning that it is strongly characterised by a specific descriptor. This four-value scale has proven successful in a number of experiments (Forsyth and Sharoff, 2014) and was deemed an acceptable trade-off between precision and confidence for annotation. In order to get cleaner text types for training purposes, pages containing two or more text types in separate areas were split into different texts. On the other hand, proper hybrid pages, i.e. those fulfilling multiple functions simultaneously, were given a strong value in each applicable attribute. This resulted in a training corpus of 931

[3] https://www.cam.ac.uk/public-engagement/about-us

[4] http://www.sci.u-szeged.hu/english/brief-history/about-us

[5] http://wwwf.imperial.ac.uk/business-school/

[6] Given that the corpus includes 91 universities, there should be at least 910 pages to code. However, two universities comprise less than 10 pages overall. Specifically, University of Rome Tor Vergata in Italy and University of Innsbruck in Austria contain two and five pages respectively.

Country	Language Family	Tokens	Texts
Germany	Germanic	1,269,884	2,674
Switzerland	Germanic-Romance	807,456	1,845
Netherlands	Germanic	801,244	1,767
Denmark	Germanic	779,139	1,382
Finland	Uralic	771,860	1,263
Sweden	Germanic	680,928	1,258
France	Romance	633,523	1,155
Italy	Romance	620,940	1,059
Spain	Romance	603,882	941
Russia	Slavic	530,522	722
Belgium	Germanic-Romance	408,088	657
Norway	Germanic	283,059	554
Austria	Germanic	185,224	352
Czech Republic	Slavic	183,370	324
Estonia	Uralic	176,162	299
Portugal	Romance	117,919	234
Slovenia	Slavic	95,309	161
Latvia	Baltic	72,568	123
Poland	Slavic	63,443	111
Romania	Romance	58,915	111
Hungary	Uralic	55,437	96
Belarus	Slavic	46,291	83
Serbia	Slavic	40,606	81
Lithuania	Baltic	36,552	44
Ukraine	Slavic	30,632	39
Greece	Hellenic	14,881	30
Slovakia	Slavic	7,905	18

Table 2: Corpus statistics by country and language family (ELF countries only).

texts. Drawing on experience from earlier annotation experiments, this number is sufficiently large to contain a representative picture of variation in academic webpages.

The annotation process produced a numeric data matrix in which each row corresponds to an observation and each column corresponds to a functional descriptor. Many texts score on several dimensions. Legal and instructional texts tend to be more recognizable, whereas informative, promotional and narrative pages show a higher degree of overlapping. Texts dealing with academic research very often score on the hardnews dimension as well, since they are often presented in the form of news bites.

The annotation matrix is used to retrieve a set of positive and negative examples for each FTD, to be used as a training set for experimenting automatic classification of the entire corpus. The amount of the positive examples for each FTD in the training corpus is listed in Table 3.

4 Automatic genre classification

Classification of texts according to their genres can be achieved by extracting a range of higher-level features, such as combinations of POS tags, parse trees or rhetorical relations (Santini et al., 2010). However, lower-level features based on character n-grams offer a surprisingly efficient method for detecting genres without requiring heavy linguistic resources (Kanaris and Stamatatos, 2007). In a comparative evaluation, their performance can exceed what is achieved by resource-heavier approaches. For example, pure n-grams can successfully generalise dates (*.*day* for *yesterday, today, Friday*), which are typical in reporting, nominalisations (*.*tion*) or passives (*.*ed by*), which are typical in scientific discourse (Sharoff et al., 2010).

The frequencies of character n-grams can be directly used as features in algorithms of Machine Learning. However, many classification methods use kernels as a mechanism for comparing the similarity between objects described by the features in order to build a model separating their classes. String Kernels (Lodhi et al., 2002) is one of such methods, which measures the similarity between webpages represented as the distance between their character n-grams.

In this study, we experimented with classification using Support Vector Machines (Smola and Schölkopf, 2004) or Relevance Vector Machines (Tipping, 2001). The advantage of RVM is the ability to produce a small number of Support Vectors, leading to better learning generalisation in the case of relatively sparse data, for example, only 25 positive examples have been identified for **A9** (legal texts). The task is to predict whether a webpage features strongly in each FTD. The commonly used F1 measure is reported in Table 3 with cross-validation for detecting the FTDs.

Once we produced reliable classifiers for each dimension, we applied them to the entire corpus of academic webpages. To establish which pages score on each dimension with minimal noise outside the training set, we experimented with reliable thresholds to achieve the desired precision. Table 3 shows the composition of the corpus in terms of the number of pages for which the predicted score is greater than or equal to each threshold and the corresponding percentage in the final corpus as opposed to the manually annotated training corpus described in Section 3.

On the whole, post-hoc evaluation shows that classification by n-grams is highly efficient in terms of precision, considering that at least 80% of pages above the threshold perfectly or widely match each specific dimension. Note that the proportion of pages that score on one dimension exclusively is very close to the one obtained from manual annotation, except for A16 dimension. The latter, however, diverges from other dimensions in that any university webpage tends to contain some degree of informational content, which may lead this dimension to be considered as a ‘safety margin’ and, eventually, to be over-represented in human annotation. Overall, approximately 50% of pages in the training set and 40% in the final corpus were classified as scoring high on one function, which is an encouraging result if we consider that online content is increasingly evolving, producing new genres and hybrid pages (Santini, 2007; Bruce, 2011).

5 Differences between language varieties

We also calculated the relative frequencies of pages that score above each threshold in order to assess their distribution across language varieties (ELF and native English) and language families (as registered in our metadata). Native English and ELF texts are equally distributed over all dimensions, apart from A16, which seems slightly more typical of ELF texts.

Looking more closely at the distribution of texts by language family (Figure 1) at least one aspect becomes immediately clear. Instructional (A7) and promotional (A12) functions are the only ones showing a medium-to-high number of pages; moreover, promotional texts are detected even in those countries that include very few pages in the original corpus, such as the Baltic and Hellenic ones, counting 144 and 30 texts respectively (Table 2). Although this may be partly related to automatic selection of pages during crawling and post-processing, the high relative frequency of promotional pages may suggest that when it comes to providing contents in English language, promotional texts are given priority over plain information, and in some cases, over instructional pages as well. A12 texts comprise very typical promotional genres, such as the ones already mentioned above (*Why choose us* pages, *About us* pages, mission statements, *Welcome* pages), as well as other texts belonging to various website sections, for instance research projects, visiting students and international strategies, descriptions of university facilities and departments, student life, sport and many others. Hard-news pages (A8) are also spread over the majority of language families, whereas legal texts (A9) appear to be relatively rare. Legal pages are slightly more frequent in Ireland and the UK where they tend to be associated with privacy policies.[7] Moving on to the A16 dimension, i.e. plain information, Romance languages seem to be separated from ELF Germanic, ELF Germanic-Romance and native English texts;[8] the former are placed between the second and fourth quartile, whereas the latter are spread below the second quartile. Greece does not include any informational pages, while Uralic and Slavic coun-

[7] http://www.ucc.ie/en/ocla/comp/data/dataprotection/

[8] Native English texts are of Germanic origin as well.

	A7	A8	A9	A12	A14	A16	A21
% in training set	8.4	5.0	3.2	8.5	6.3	13.6	5.5
F-measure	0.95	0.92	0.96	0.85	0.93	0.79	0.94
% in final 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

Table 3: Manual annotation of the training set and final corpus.

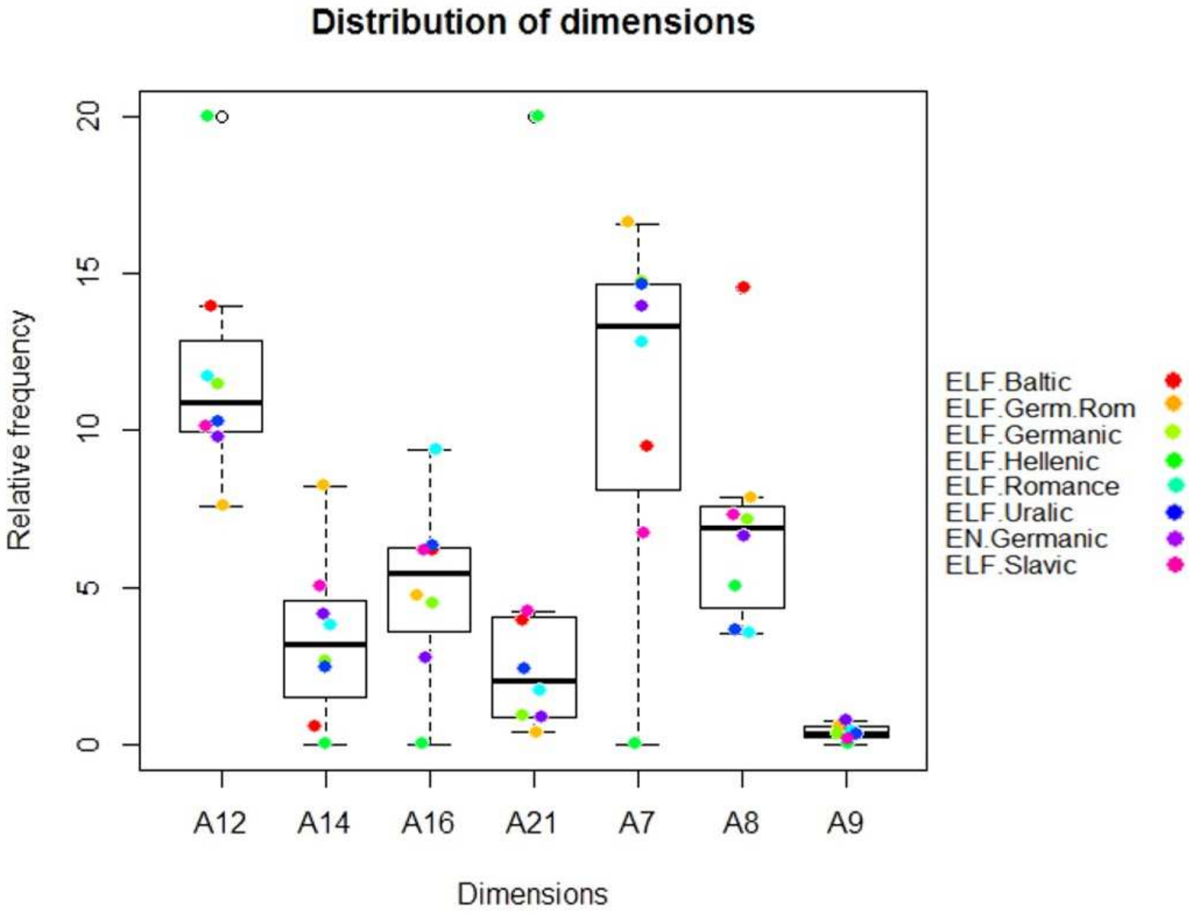

Figure 1: Distribution of texts by language variety and language family.

tries are closer to the Romance ones. Examples of informational texts include lists of items,[9] descriptions of university services and administrative offices.[10]

Pages reporting academic research (A14) are less evenly distributed. Switzerland is the country with the highest number of texts representing academic research, whereas Hellenic and Baltic countries have next to no pages in the corpus on this dimension. Finally, narrative texts - i.e. pages describing chronologically ordered events - place themselves between legal and research pages, showing higher frequencies in Slavic, Uralic and Baltic regions, and a very high peak in Greece. Genres from this dimension include university history in Greece,[11] the description of historical figures in Romania,[12] Professors academic careers in Ukraine and the description of university museums in Estonia.[13]

By exploiting URL strings, one can also detect typical website sections in order to analyse a) how language is used in the same dimensions across English varieties and families and b) how language is used across different dimensions. For instance, when searching the string *why* among pages that score highly on the A12 dimension, 78 texts are retrieved overall, each of them matching the genre *Why choose us*. Although no systematic analysis of language features has been performed yet, some interesting patterns emerge when analysing these pages by language variety. Besides native English and ELF dissimilarities that have already been observed in previous studies (Bernardini et al., 2010) - e.g. a larger use of second person pronouns by native English universities - from the point of view of content, *Why choose us* texts produced in Ireland and the UK make more frequent references to *help* and *support*, as compared to ELF pages. On the other hand, in ELF texts there is repeated mention of the *international* and *European* perspective that seems to be less common among native English countries. As far as the second type of analysis is concerned, searching the string *mission* among texts that score highly on A16 and A12 dimensions will yield two completely different text types. Example 1 and Example 2 below are two excerpts of mission statements taken, respectively, from the University of Vienna[14] and from Imperial College London.[15] As predicted by automatic classification, Example 1 scores highly on the A16 dimension, whereas Example 2 scores on the A12 dimension.

(1) The International Office serves as an information hub and service facility in the field of internationalisation and international relations at the University of Vienna. We support and advise members of the university in all international agendas, in particular in relation to requests for bilateral cooperation projects. The International Office is also involved in the implementation of the internationalization strategy of the University of Vienna.

(2) The Graduate School plays a key role in delivering the postgraduate student experience as well as with postgraduate education, policy and strategy development. The Graduate School enriches the postgraduate student experience by delivering a tailored programme of professional skills training which enhances the professional impact and helps to ensure personal ambitions are realised.

Although both texts are placed on the same website section named *mission* or *our mission*, from an internal perspective they are different. Example 1 adopts language patterns that usually characterise administrative texts (*serves as, in relation to requests, implementation of*), whereas Example 2 employs positive loaded words that are very typical of evaluative language (*key, enrich, enhance, ambitions realised*) and mission statements as well (Morrish and Sauntson, 2013). Besides confirming the performance of classification based on n-grams, these two examples raise some issues related to the efficiency of reflexive categories (Sinclair and Ball, 1996), especially when university webpage titles refer to genre, rather than topic.

[9] http://www.bsu.by/en/main.aspx?guid=134021

[10] http://www.unibo.it/en/university/campuses-and-structures/urp-public-relations-office/services-urp

[11] http://www.ntua.gr/history_en.html

[12] http://150.uaic.ro/personalitati/biologie/ioan-borcea/?lang=en

[13] http://www.univ.kiev.ua/en/geninf/adm/Zacusilo/

[14] http://zid.univie.ac.at/en/about-us/vision-mission/

[15] https://www.imperial.ac.uk/study/pg/graduate-school/about-us/mission-statement-/

6 Conclusions and further research

This paper reports an experiment on automatic classification and analysis of a corpus of university webpages in terms of genres by using string kernels with the aim of exploring the distribution of genres across English varieties and English language families. Classification by n-grams has proven successful in terms of precision. Post-hoc evaluation showed that more than 80% of pages above the reliability thresholds match the predicted dimension.

Instructional and promotional webpages have the largest share in our corpus across all language varieties, such as English native and ELF. However, variation is higher when considering each language family. In a few cases, variation may be related to country-specific aspects and how universities wish to present themsleves internationally, for instance Greece focusing on university history and Switzerland showing the highest number of texts related to academic research. Universities located in a country where the official language is of Romance origin exhibit the highest number of plain information, partially due to the descriptions of university offices and services. The informational dimension seems to be quite uncommon in ELF-Germanic and Native English texts, where it reaches its lowest levels, i.e. Ireland, the UK, Belgium and Denmark.

Automatic classification of university web genres enables comparison of genres across dimensions and language varieties. Although findings have not been generalised to the full set of our data, they form the basis for future systematic analysis across text types, genres and English language varieties in university websites. In the future, we plan to carry out clustering to identify hybrid texts and genre categories that score on more than one functional dimension simultaneously, such as info-promotional pages and news describing academic research. Other plans include investigating the relation between text types and other linguistic or contextual information, such as university world ranking. Finally, this work also carries applied implications for developing and improving communicative strategies based on the analysis of typical features of highly-ranked universities, as suggested by the examples provided at the end of the previous section.

Acknowledgments

We are grateful to Silvia Bernardini for her extensive comments on the earlier drafts.

References

Inger Askehave. 2007. The impact of marketization on higher education genres: The international student prospectus as a case in point. *Discourse Studies*, 9:723–742.

Marco Baroni and Silvia Bernardini, editors. 2006. *Wacky! Working papers on the Web as Corpus*. GEDIT, Bologna.

Silvia Bernardini and Adriano Ferraresi. 2013. The academic Web-as-Corpus. In *Proceedings of the 8th Web as Corpus Workshop*, Lancaster.

Silvia Bernardini, Adriano Ferraresi, and Federico Gaspari. 2010. Institutional academic English in the European context: a web-as-corpus approach to comparing native and non-native language. In *Professional English in the European context: The EHEA challenge*, pages 27 – 53. Peter Lang, BERNA.

Ian Bruce. 2011. Evolving genres in online domains: the hybrid genre of the participatory news article. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Text, Speech and Language Technology*. Springer.

Luisa Caiazzo. 2011. Hybridization in institutional language: Exploring we in the 'about us' page of university websites. In Srikant Sarangi, Vanda Polese, and Giuditta Caliendo, editors, *Genre(s) on the move. Hybridization and discourse change in specialized communication*, pages 243–260. Edizioni Scientifiche Italiane, Napoli.

Alberto Fernndez Costales. 2012. The internationalization of institutional websites: The case of universities in the European Union. In A. Pym and D. Orrego-Carmona, editors, *Translation Reasearch Projects 4*. Intercultural Studies Group, Terragona.

Erika Dalan. 2015. Classifying university websites: A case study. Poster session presented at Corpus Linguistics 2015, Lancaster, July.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*.

Richard Forsyth and Serge Sharoff. 2014. Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, 29:6–22.

Sara Gesuato. 2011. Course descriptions: Communicative practices of an institutional genre.

In Srikant Sarangi, Vanda Polese, and Giuditta Caliendo, editors, *Genre(s) on the move. Hybridization and discourse change in specialized communication*, pages 221–241. Edizioni Scientifiche Italiane, Napoli.

Ioannis Kanaris and Efstathios Stamatatos. 2007. Webpage genre identification using variable-length character n-grams. In *Proceedings of ICTAI*.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444.

Sandra Mollin. 2006. English as a lingua franca: A new variety in the new expanding circle? *Nordic Journal of English Studies*, 5(1):41–57.

Liz Morrish and Helen Sauntson. 2013. Business-facing motors for economic development: an appraisal analysis of visions and values in the marketised uk university. *Critical Discourse Studies*, 10(1):61–80.

Giuseppe Palumbo. 2013. Divided loyalties? Some notes on translating university websites into English. *CULTUS*, 6:95–109.

Kem Saichaie. 2011. *Representation on college and university websites: An approach using critical discourse analysis*. Ph.D. thesis, University of Iowa.

Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. In Alexander Mehler, Serge Sharoff, and Marina Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Marina Santini. 2007. Characterizing genres of web pages: Genre hybridism and individualization. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pages 71 – 81, Washington, DC, USA. IEEE Computer Society.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. In *Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010*, Malta.

Serge Sharoff. 2015. Approaching genre classification via syndromes. In *Proc Corpus Linguistics*, Lancaster.

John Sinclair and Jackie Ball. 1996. Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document.

Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222.

Michael E Tipping. 2001. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.