

Dialects of Nordic English

Alex Tang, Liam Sullivan

School of Computing, Leeds University, UK

Potential Journal: Nordic Journal of Linguistics

sc07at@leeds.ac.uk

sc07ls@leeds.ac.uk

Introduction

Dialects of English are used throughout the world, our interest for this research has been to classify English into two distinct categories, British English and American English. Research communities almost exclusively favour English for research published to an audience of readers with mixed mother tongues, state and litigation proceedings are often set in English even in some cases where it is not the official language of the country. An example of just how popular our chosen categories are is evident when one considers that most journal editorial boards require authors to present research papers in one of the two, and will reject English of a mixed dialect.¹

The distinction between these two very widely used dialects of English, American and British, is frequently identified only by a small subset of popular words with different spellings across the two. Our interests for this paper are to classify the language spoken in our chosen collection of Nordic countries, specifically; Greenland, Iceland, Denmark, Norway and Sweden. We improve on past research by extracting objective, empirical evidence to support our opinions on which type of English more closely models that spoken in these countries, and the region as a whole.

Before applying our practical data-mining techniques we first needed to collate samples of language from English websites based in the region. These corpora were extracted from the web with each set being constrained using the Top Level Domains listed in **Table 1**, to limit results to those territories. These samples were used in a comparison with the WWW corpus for the United Kingdom and U.S.A.

Table 1

Country	Top Level Domain
Greenland	.gl

¹Requirements for contributions to Journal: Nordic Journal of Linguistics http://assets.cambridge.org/NJL/NJL_ifc.pdf

Iceland	.is
Denmark	.dk
Norway	.no
Sweden	.se

To build on the successes of others embarking on research in this area we chose to follow the CRISP-DM Data-Mining methodology (Chapman et al 2000). Classification evidence was gathered through use of the WEKA Data-Mining toolkit (Witten and Frank 2005). The data we obtained, accompanied by possible explanations for the results follows.

It goes without saying that English is a truly global language, in all of the countries studied a majority of the population speak at least moderate amounts of the language. The authors are not aware of any formal research specifically asking the question of which type of English has the most influence on the dialect in use within these countries.

The authors expect this paper to be of interest to researchers with an active interest in corpus data-mining approach to language analysis, those considering including techniques in their future projects, and those working with Nordic languages. For example the Nordic Journal of Linguistics (NJL, Cambridge University Press) contains contributions of “general and methodological interest and studies on Nordic languages” , and describes itself as a “major forum for linguist working in a Nordic country”, the authors expect the study will be of interest to the NJL’s readership, notably due to the connections with the most recent volume containing a report on Nordic citizens learning second languages.

Investigations into the use of language within a community or geographical region often include subjective opinion based suggestions which can skew results. We argue that our data-mining approach provides a true top down view on the use of a language within a given region, therefore it is the authors belief that data-mining has a lot to offer language researchers. As the field of Computational linguistics grows the availability of user-friendly data-mining and language processing software, making the area accessible to those linguists and language experts outside the computer science sphere. This interdisciplinary area is rapidly expanding and more linguists are turning to computational models over traditional methods, we hope that our study will offer an insight to those in the readership who are yet to include classification tools such as WEKA.

During the initial stages of the study we discussed our predictions for the outcome, which dialect of English is likely to dominate in the Nordic region, and each of its individual countries. The most obvious factor to consider is geographic distance, on face value a fair

informed guess could be that Greenland and Iceland tend towards an American version, whilst the remaining countries exhibit a majority in British English due to the close uninterrupted connection to the UK across mainland Europe, with the North Atlantic Ocean creating a natural distinction between the UK and Greenland.

Another potential factor is EU membership; Norway and Iceland are not part of the European Union and this absence of British English for litigation and official EU business could redress the balance between the teaching of English with motivations relating to business or education/general interest. After debating each factor we recorded some solid predictions, settling on a high likelihood of Iceland and Greenland tending towards American, and the remaining Nordic regions following a majority in British English due to the issues discussed above. Whilst it could be argued that the geographical factors would imply a horizontal divide on the map of each country as the potential influence of British English declined, for this is disregarded for the purposes of our study as we are considering each territory as a whole. Following our initial predictions we moved onto the characteristic decision making process.

Methods

To create a pool of representative text for analysis we used WWW corpora consisting of 200,000 web pages for each country; Norway, Iceland, Sweden, Greenland and Finland. The corpora contain language gathered from web pages, with each individual corpus containing only content gathered from the respective top level domain as given in **Table 1**. These representative texts were compared with a UK and US English text samples of the same volume.

The World Wide Web includes a Country Code Top Level Domain system regulated by ICANN (Internet Corporation for Assigned Names and Numbers), this allows restrictions to be imposed at the sample gathering stage simply and consistently by only considering pages ending in the ccTLD in question for each case. The samples we used in this study were obtained by students in the School of Computing at University of Leeds. The tool of choice was Google, in particular the Advanced Search options that make it possible to restrict results to a specified domain and language.

An additional tool, WWW-BootCat (Baroni et al 2006) a web-based interface to BootCat (Baroni and Bernardini 2004) was used on top of Google, with communication through the Google API providing a powerful combination for gathering the corpus. This interface uses seeds words, typical English words (Sharoff 2006). The final stage of the corpus gathering is made easier by another useful feature provided by Bootcat, that of 'cleaning' the content removing HTML and other mark-up associated with instructing the web browser on how to present the text, leaving just the English text. Although the web is by no means exclusively English the fact remains that a large enough proportion of each territory we analysed allowed

the extraction of a representative sample.

During this stage our goal was to find the differentiating features most appropriate to use as classifiers to decide which type of English the sample most closely matches to. The most obvious differentiating features are words spelt differently in each version of English, such as color and colour for example. After researching other variations such as placement of verbs within a sentence and also the existence of country specific mentions eg. BBC or President., we decided upon a set of these words that are spelt differently. Through experimenting with the classifiers in Weka we decided to include tokens for both spellings and also a figure for percentage out of total for each, as shown in **Table 2**.

Table 2

British English Word	American English Word
centre	center
colour	color
organisation	organization
program	programme

Because we needed to focus on the relative frequency of the British English words against their US counterparts the sample text needed to be processed to output a Word Frequency List. This easy to realise with a few basic Linux commands strung together, shown below (Atwell)

```
sort corpus | uniq -c | sort -n -r > wordFreqList
```

The final stage of the process before obtaining classification results from Weka is to provide the data-mining software with these word frequencies, in the form of an arff file as show in **Code Sample 1**. A similar arff files was also constructed to build the training set from our example data of standard UK and US text, this was then used to train the classifiers in Weka. The **@attribute** tag is used to record characteristics we would like Weka to use in classification, with these values being recorded as numerical values below based on the results of word frequency analysis for each country's sample text. It should be noted that the software may not use all values after initial training to find the "best fit" model for the algorithm.

Code Sample 1

```
@relation nordic
```

```
@attribute center numeric
```

@attribute centre numeric
@attribute centerpercent numeric
@attribute centrepercent numeric

@attribute color numeric
@attribute colour numeric
@attribute colorpercent numeric
@attribute colourpercent numeric

@attribute organization numeric
@attribute organisation numeric
@attribute organizationpercent numeric
@attribute organisationpercent numeric

@attribute program numeric
@attribute programme numeric
@attribute programpercent numeric
@attribute programmepercent numeric

@attribute english {UK,US}

@data

100, 84, 54, 46, 53, 7, 88, 12, 19, 30, 39, 61, 142, 72, 66, 34, UK
91, 47, 66, 34, 4, 7, 36, 64, 39, 6, 87, 13, 156, 35, 82, 18, UK
20, 32, 38, 62, 39, 26, 60, 40, 13, 8, 62, 38, 49, 10, 83, 17, UK
13, 24, 48, 52, 13, 8, 62, 38, 9, 13, 40, 60, 58, 25, 70, 30, UK
43, 34, 56, 44, 12, 27, 31, 69, 41, 24, 63, 27, 114, 35, 77, 33, UK

Results

From the chosen key features, we improved the given training data for the UK and US training samples by extracting the number of times they occurred in each frequency list for the 10 UK and US samples. We continued to separately extract the equivalent frequencies for our Nordic region English web-corpora, to use the WEKA toolkit in testing the model and classifying them as either UK or US English.

The manually selected features taken from the web-corpora were not always used, when loaded into the WEKA toolkit only one feature was used to classify the Nordic samples as either UK or US. This feature changed based on which classifier was used, an overview of the features are shown in Figure 1 below.

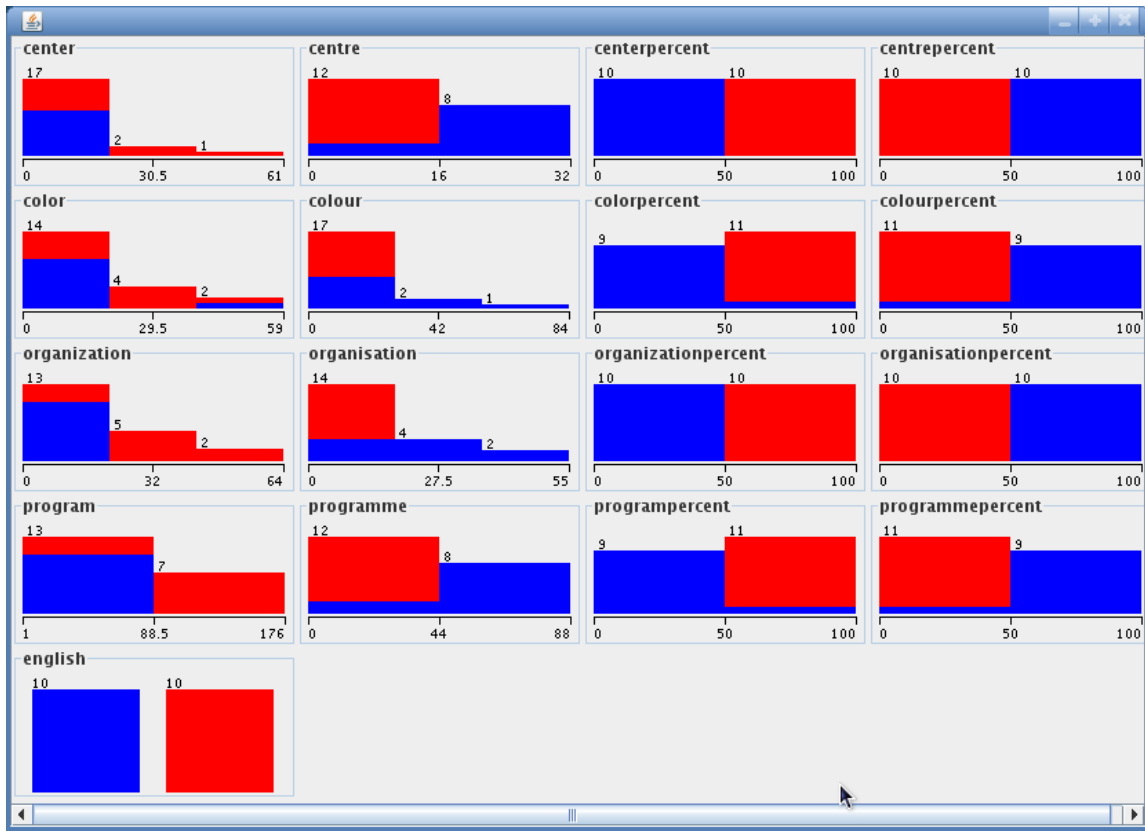


Figure 1: Visualisation of all the features in the training data set.

The JRip and OneR rule based classifiers used the centre feature as its model to classify our samples. This resulted in correctly classifying all 5 instances as UK giving 100% accuracy, the confusion matrix from WEKA is shown below.

```
=== Confusion Matrix ===
 a b <-- classified as
 5 0 | a = UK
 0 0 | b = US
```

The result was then enforced by the ZeroR rule based classifiers, by also correctly classifying all 5 instances as UK with 100% accuracy and therefore producing the same confusion matrix.

Using the J48 decision tree classifier selected the feature organizationpercent to classify whether our samples were either UK or US.

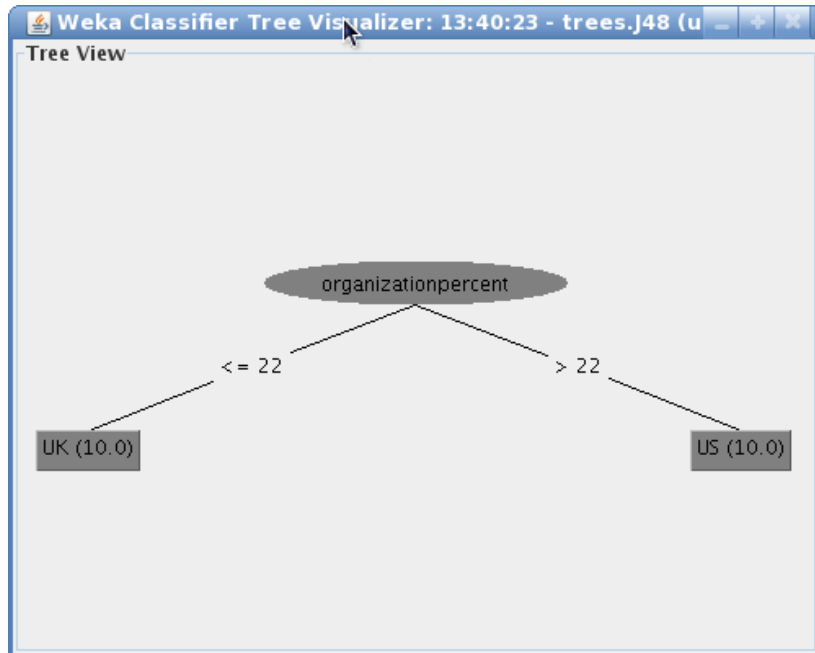


Figure 2: Decision tree from the J48 classifier in WEKA.

This classified all 5 of our samples were more like US English compared to our initial prediction of them being more like UK English which led to the confusion matrix below.

```
=== Confusion Matrix ===
a b <-- classified as
0 5 | a = UK
0 0 | b = US
```

Contradicting the earlier results using rule based classifiers by 100% incorrectly classifying all 5 instances, we ran another decision tree classifier.

The NativeBayes decision tree classifier resulted in a 80% accuracy in correctly classifying 4 samples as UK English. Predicting Greenland being more like US English, the raw data shows that the organizationpercent had 87% for this sample explaining why this result came about.

All WEKA outputs for the classifiers are included in the Appendix where the rules and confusion matrices can be compared.

For the overall result of the 5 Nordic countries, we found:

English in .dk .gl .is .no .se (Denmark, Greenland, Iceland, Norway, Sweden) is more like .uk English. However this is not 100% accurate based on all of our Data Mining tools used, with one tending all samples being more like US English.

We then combined all 8 national English samples into a single .NORDIC Nordic English corpus, and used the provided corpus-comparison software tool `compare-fq-lists.pl` to compare .NORDIC against UK and US English standards.

We used the Log-Likelihood corpus comparison tool to find words which were more frequent in Nordic English web-text than in British English web-text. The words with most significant differences in frequency are shown below.

Word	Frq1	Frq2	LL-score	
s	65	4266	6310	
t	22	977	1397	
Arctic	49	948	1202	
Iceland	11	715	1057	
Greenland	10	631	931	
upload		5	476	721
listings		11	480	685
Icelandic	10	463	665	
don	6	391	578	
Danish		25	401	489

Names of places locally significant to Nordic countries , for example: Arctic, Iceland, Greenland, were often the words with high Log-Likelihood scores however this is not really a feature of linguistic.

s and t : Nordic web-pages use much more contractions, for example *don't v do not*, *can't v can not* placing these in the top 2 findings.

To look at the opposite end of the frequency differences we used the Log-Likelihood corpus comparison tool to find words which are relatively more frequent in British English compared to Nordic English. The words with most significant differences in frequency are shown below.

Word	Frq1	Frq2	LL-score	
a	16172	33154	2817	
I	4306	10532	1500	
BBC	15	1349	1448	
you	3942	9646	1376	
for	10378	19943	1326	
to	27496	45137	1315	
UK	172	1542	977	
London	105	1176	828	
Wales		6	586	633
aircraft		18	641	619

The words with high Log-Likelihood scores seem to be names of places locally

significant again. For example: BBC, UK, London are more common in British English web-texts.

I, you, for and to shows the way of interaction between people in British English web-texts, placing them in the top findings.

Conclusions

From our collection of Data Mining analyses, the evidences concludes that the status of English used in each of the Nordic countries predominately indicate:

English in .dk .gl .is .no .se (Denmark, Greenland, Iceland, Norway, Sweden) is more like .uk English. However this is not 100% accurate based on all of our Data Mining tools used, with one tending all samples being more like US English.

This reflects on our initial assumption of Iceland and Greenland tending towards American, and the remaining Nordic regions following a majority in British English being different. However based on different features used for classification we saw that all 5 of the national samples seemed to be more like US English with one of the Data Mining tool. This could reflect on our chosen features using to determine UK English from US English showing a limitation to our approach.

The result for Greenland is most significant, as due to its geographic location it appears to favour American English. One reason maybe with a large amount of scientists working in that area being American many web papers can be seen to be more like US English.

Overall, we can say American nor British English is truly dominant across the Nordic region based on our evidence, but there is strong tendency to prefer British English around that area. Due to the countries lying within Europe it can be expected that they will tend to be more like British English, however with studies in the colder regions of Greenland taking place by a majority of Americans it is plausible to see that country tend towards US English.

Finally, as an additional exercise, we combined all 8 national English samples into a single .NORDIC Nordic English corpus, and used the provided corpus-comparison software tool `compare-fq-lists.pl` to compare .NORDIC against UK and US English standards. A web-text corpus can only give us lexical differences; differences in accent or pronunciation don't show up on WWW texts.

Names of places locally significant were the words with the highest Log-Likelihood scores however this is not really a feature of linguistic. For example: Arctic, Iceland,

Greenland for Nordic countries whilst BBC, UK, London are more common in British English web-texts.

We found no clear overall preference for British v American spelling, eg *color* v *colour*, *centre* v *center* – this is consistent with our overall conclusion that neither UK nor US English dominates in the Nordic region.

References

- Atwell, E; Arshad, J; Lai, C; Nim, L; Rezapour Ashregi, N; Wang, J; Washtell, J. 2007. Which English dominates the World Wide Web, British or American? In: Proceedings of Corpus Linguistics 2007, Birmingham.
- Atwell, E; Abu Shawar, B. 2008. An AI-inspired intelligent agent/student architecture to combine language resources research and teaching. In: Proceedings of LREC'08: Language Resources and Evaluation Conference, Marrakech.
- Baroni, Marco and Bernardini, Silvia. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon, pp 1313-1316
- Baroni, Marco; Kilgarriff, Adam; Pomikalek, Jan; Rychly, Pavel. 2006. WebBootCaT: instant domain-specific corpora to support human translators. In: Proceedings of EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation., pp 247-252.
- Bickerton, Anthea. 1971. American English / English American: a two-way glossary of words in daily use on both sides of the Atlantic. London: Abson Books.
- Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin and Wirth, Rudiger. 2000. CRISP-DM 1.0 Step-by-step data mining guide. <http://www.crisp-dm.org/CRISPWP-0800.pdf>
- Crystal, David 2003. English as a Global Language (second edition). Cambridge University Press.
- Dakroury, A. 2008. The Arab-Canadian consumption of diasporic media. International Journal of Communication, volume 12.2
(N.J.L, Cambridge University Press)
http://assets.cambridge.org/NJL/NJL_ifc.pdf
- Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. In: M. Baroni, S. Bernardini (eds.) WaCky! Working papers on the Web as Corpus, Bologna. <http://corpus.leeds.ac.uk/serge/publications/wacky-paper.pdf>
- Sharoff, S. 2007. Classifying Web corpora into domain and genre using automatic feature identification. In Proceedings of Web as Corpus Workshop, Louvain-la-

Neuve, September 2007.

Trudgill, Peter and Hannah, Jean. 2008. International English: A Guide to the varieties of Standard English (fifth edition), Hodder Education.

Witten, Ian and Frank, Eibe. 2005. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann.

Appendix