
STUDIES IN VARIATION, CONTACTS AND CHANGE IN ENGLISH

VOLUME 15

Texts and Discourses of New Media

Choice of national variety in the English-language Wikipedia

Turo Hiltunen

University of Helsinki

Please cite this article as:

Hiltunen, Turo. 2014. “Choice of national variety in the English-language Wikipedia”. *Texts and Discourses of New Media* (Studies in Variation, Contacts and Change in English 15), ed. by Jukka Tyrkkö & Sirpa Leppänen. Helsinki: VARIENG.
https://urn.fi/URN:NBN:fi:varieng:series-15-4

BibTeX format

```
@incollection{Hiltunen2014,
  author = "Turo Hiltunen",
  title = "Choice of national variety in the English-language Wikipedia",
  series = "Studies in Variation, Contacts and Change in English",
  year = 2014,
  booktitle = "Texts and Discourses of New Media",
  number = "15",
  editor = "Tyrkkö, Jukka and Leppänen, Sirpa",
  publisher = "VARIENG",
  address = "Helsinki",
  url = "https://urn.fi/URN:NBN:fi:varieng:series-15-4",
  issn = "1797-4453"
}
```

Abstract

The English-language Wikipedia has been widely used in computational linguistic studies, but there is less corpus-based work on the language of Wikipedia from a variationist perspective. This article investigates whether Wikipedia articles follow American English or British English usage norms to a different degree, and to what extent this depends on their subject matter. The data for this study comes from *Westbury Lab Wikipedia Corpus (2010)*. The results indicate that Wikipedia articles in general favour American English usage norms, which reflects the large number of contributing editors based in the US. On the other hand, British English holds its own in articles with ties to the UK, which exhibit a higher incidence of spellings and grammatical forms associated with that variety.

1. Introduction

In recent years, Wikipedia has become so widespread that it hardly requires an introduction. According to Wikipedia's entry on Wikipedia itself, it is a 'collaboratively edited, multilingual, free Internet encyclopedia' ([[Wikipedia]]). [1] Established in 2001, it has become one of the most popular pages on the Internet, with 365 million readers and over 12 billion monthly page views globally. [2] It has been hailed as one of 'Internet's most radical successes during the first decade of the twenty-first century' (Lih 2009: 225) – this quote comes from the Afterword to Andrew Lih's book *The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia*, which has itself been written via an online wiki. [3] This book also contains a Foreword written by Jimmy Wales (co-founder of Wikipedia), which begins with the following description of Wikipedia's mission:

> Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing. (Lih 2009: xvi)

A key feature of Wikipedia is that unlike a traditional encyclopedia, it can be edited by anyone with Internet access. Every contributor to Wikipedia is referred to as an 'editor'. Some editors use their real name, but most editions are made anonymoysly, which obviously opens up a variety of questions regarding reliability and credit (see e.g. Hoffmann 2008). The content of articles is regulated by a number of policies ([[List of Policies]]) and guidelines ([[List of Guidelines]]), which editors should be familiar with. For example, Wikipedia's three core content policies are [[Neutral point

of view]], [[Verifiability]], and [[No original research]], which state that the articles should represent all relevant views in an unbiased fashion and that all claims should be attributed to reliable secondary sources. The guidelines, meanwhile, describe the preferred forms of interaction between editors and provide guidance on matters of style.

This article is concerned with the question of how data from Wikipedia can be used in corpus linguistic research. Being a freely available source of large amounts of text, Wikipedia has obvious potential for exploitation in various kinds of linguistic applications. In fact, data from Wikipedia has already been used extensively by computer scientists and computational linguists. In a recent review article, Medelyan et al. (2009) identify four categories of this work: natural language processing, information retrieval, information extraction and ontology building. At the same time, despite the growing popularity of web-derived corpora in general, Wikipedia has so far been relatively little used in corpus linguistic research.

There may be various reasons for the slow adoption of Wikipedia data in corpus linguistics. For one, given the enormity of Wikipedia as a resource, it is not obvious how it can be converted into a format that is useful for linguistic research. And if the aim is to use Wikipedia as a corpus (the 'Web for corpus building' approach, see Section 2), it is also necessary to take account of its dynamic nature. Typically, written language corpora contain texts that have already been published and will not be modified at a later stage. However, with Wikipedia this is not possible by definition, because articles may be edited any time after they have been downloaded by the corpus compiler (see also Section 6.4).

Within applied linguistics, Wikipedia has similarly attracted less scholarly attention than other, more established genres. This is not surprising altogether: the role of Wikipedia is often considered to be problematic in educational settings. It is frequently criticised for, among other things, inaccuracy, triviality and bias. [4] It is commonly agreed that like any other tertiary source, Wikipedia should only be used as a 'first port of call' in a research project, and consequently not be cited in essays or papers (see e.g. Walters 2007). [5] For this reason, it is not surprising that Wikipedia articles would hold less interest for applied linguists than other genres which are academically and professionally more relevant to students.

From another perspective, however, the question of how language is used in Wikipedia is clearly relevant to many areas of applied linguistics, including teaching of English for academic purposes (EAP). First, the language of Wikipedia articles is specialised variety of English to which students are frequently exposed; there is ample evidence that the use of Wikipedia is extremely widespread among both students and academics

(Eijkman 2010, Konieczny 2014). In addition, Wikipedia has been successfully used in collaborative learning projects in a variety of circumstances. For example, Matt Barton, assistant professor of English at St. Cloud State University, Minnesota, has collaborated with his students to create resource for teaching rhetoric in first-year college composition programs (Barton 2005, Barton and Cummings (eds.) 2008), which makes extensive use of Wikipedia. MacLeod (2007) reports on a course at the University of East Anglia, where students write and edit articles on controversial topics in Middle Eastern politics. According to Tapscott and Williams (2010: 150), use of wikis may help students take an active role in the construction of knowledge, which has been shown to be an effective method of learning. This being the case, it is clear that Wikipedia can provide useful and valuable information for linguistic research beyond the areas of computational linguistics and natural language processing.

The primary aim of this study is to show that despite the possible limitations discussed above, Wikipedia can be used as an resource for analysing linguistic variation from a corpus linguistic perspective, and that it can provide useful information about determinants of linguistic variability. To this end, the article presents a case study on the preferred national variety of English in Wikipedia articles. The question of variability is an interesting one in Wikipedia, because writers and editors frequently need to choose (whether consciously or not) between two or more words or grammatical structures which may be more viable in a particular regional variety. Wikipedia itself takes a neutral position on the varieties of English, with the exception of a small number of situations, including 'ties to a particular English-speaking nation' (see Section 3). For our purposes, the interesting question then is how consistently particular forms are chosen over others, and this study aims to demonstrate that the answer to this question is linked with their subject matter of the article. The article also considers the benefits and disadvantages of Wikipedia as a source of data specifically in variationist linguistics.

Accordingly, I shall investigate different samples extracted from the English Wikipedia to obtain information about the relative prominence of forms which are associated with different national varieties. The focus will be on British English and American English, the two principal national varieties of the language (Algeo 2006: 2). Specifically, the article aims to provide answers to two questions:

1. Do articles in general favour British of American spellings?
2. Do articles with ties to the UK follow British spellings, and do articles with ties to the US follow American Spellings?

The article is structured in the following way. Section 2 presents an overview of earlier work on the language of Wikipedia. Section 3 discusses the issue of AmE/BrE variation

in Wikipedia articles. Section 4 describes the data used in this study, and Section 5 presents the method of analysis employed. The results of the analysis are presented in Section 6, and their implications are discussed in Sections 7 and 8.

2. Corpus linguistics and Wikipedia

Web-derived text has gained a firm foothold in corpus linguistics in recent years (see e.g. Kilgarriff and Grefenstette 2003, Hoffmann 2007, Hundt et al. (eds.) 2007a, Kehoe and Gee 2007, Hundt 2009, and Davies 2010). The amount of freely available text offers extremely attractive perspectives for corpus linguists. At the same time, data from the Internet presents various challenges to the researcher, having to do with such questions as how to clean up messy web texts efficiently or whether search engine results actually offer reliable frequency data for linguistic phenomena.

Wikipedia is thus only one of the numerous online resources that can be used for linguistic analysis. As previously described, Wikipedia has already been used extensively in some areas of linguistic research, in particular text mining and information retrieval research (see Medelyan et al. 2009), but it has clearly been underused as a corpus linguistic resource outside the statistical NLP community (Walton 2009: 2, cf. Pentzold and Seidenglanz 2006: 59). Hundt et al. (2007b: 2) distinguish between two ways of using Internet data for corpus linguistic research: the "Web as corpus" approach, where the Internet is used directly as a corpus with the help of commercial crawlers and search engines, and the "Web for corpus building approach", where it is used or as a source for the compilation of offline monitor corpora. Earlier corpus linguistic studies using Wikipedia represent exclusively the latter approach: they involve downloading and annotating a segment of Wikipedia data and using it as an offline corpus. This is also the approach that is adopted in the present paper (see Sections 4 and 5).

Earlier studies on the English Wikipedia have mainly focussed on questions of style, readability and the nature of discourse. Particularly prominent themes include such stylistic issues as formality and homogeneity. For example, using a variation of Biber's (1988) multidimensional analysis, Emigh and Herring (2005) found that Wikipedia articles are equally formal as traditional print encyclopedia articles (taken from *Columbia Encyclopedia*), but more formal than articles in another online encyclopedia (*Everything*$_2$). For Emigh and Herring, these results show that the 'neutral point of view' principle, one of the three core content policies determining what kind of material is acceptable in Wikipedia articles, actually produces articles that resemble traditional encyclopedia articles. [6] A simplified version of this methodology is used

by Elia (2007), who compares Wikipedia to *Encyclopedia Britannica* with respect to features which Biber (1988) had identified as having high positive or negative loadings on Dimension 1, ('Informational vs. Involved production'). She found that a random sample of 100 Wikipedia articles scored slightly lower on formality than their counterparts in *Encyclopedia Britannica*. A third study with a similar orientation is Walton (2009), which studies some stylometric variables, such as word and sentence lengths. Walton observes a positive correlation between homogeneity and the number of edit counts, such that the more times articles are edited, the more homogeneous they become with respect to word and sentence length.

As for readability, Elia (2009) concludes that Wikipedia articles are comparable to popular American magazines such as *Time* and *Newsweek*. The question of readability is discussed also in Walton (2009: 40), who concludes that unlike the stylometric variables mentioned earlier, readability scores do not appear to be correlated with edit counts.

In addition, some studies take a discourse-analytical perspective on Wikipedia texts. Along wth studying the main text of the articles, these studies make use of the 'History' and 'Talk' pages, which make visible the process through which the current version of the article has emerged. For example, Pentzold and Seidenglanz (2006) apply Foucault's discourse theory to the analysis of communicative user interaction in Wikipedia, concentrating on the article [[Conspiracy theory]] and its page history over a period of four months. Based on their analysis of how authors produce and edit text and resolve conflicts, they argue that their collaboration is not chaotic but follows ordered procedures to 'delimit the sayable' (2006: 78). In a recent textbook, Myers (2010) illustrates the ways of analysing the discourse of Wikipedia. He looks at the first 100 edits of two articles, [[Manchester]] and [[7 World Trade Center]], finding considerable differences in their production history. In addition, he analyses a 5,000-word sample of the 'Talk' pages of twenty articles, showing that Wikipedians use a wide variety of rhetorical devices and frequently invoke explicit principles such as 'neutral point of view' when arguing their case (2010: 146–151).

In contrast to the aforementioned discourse-oriented studies, this paper takes a variationist approach to the language of Wikipedia, focusing on the choice of national variety in different samples extracted from Wikipedia. Variationist analysis is here understood as being concerned with finding out the extent to which alternative realisations of a specific linguistic vary systematically (e.g. Jucker 1992). While the approach was originally developed within correlational sociolinguistics (e.g. Labov 1966 and Labov et al. 1968), it has more recently become a well-established method in corpus linguistic studies on lexical and grammatical variation (see e.g. Romaine 2008 and Mair 2009b: 24–25.)

This kind of research is made easier by fact that many large corpora compiled from Wikipedia text have recently been made available. These corpora are designed with slightly different research goals in mind, which is shown in their different make-up. The *Wikipedia XML corpus* (Denoyer and Gallinari 2006), for example, preserves information about the hierarchical structure of the categories in XML format, and is therefore useful for research into the structure and interrelationship of articles. Another recently completed corpus is the *Wikicorpus* (Reese et al. 2010), which contains 750 million of words complete with lemmatisation, POS-tagging and word sense annotation. Both these corpora contain texts in a number of languages, which makes them potentially useful for contrastive analysis and translation studies. [7] The present study uses the largest Wikipedia corpus of English texts currently available, *The Westbury Lab Wikipedia Corpus (2010)* (Shaoul and Westbury 2010). It is is essentially a snapshot of all Wikipedia articles longer than than 2,000 characters. [8] The corpus contains 2 million individual articles and 990,248,478 words. Finally, *WaCkypedia_EN*, compiled as part of the *WacKy* project (Baroni et al. 2009), is similarly based on the Wikipedia database dump (a record of the entire Wikipedia database) downloaded in 2009), and it has been released with POS-tagging, lemmatisation and dependency parsing. [9]

3. British or American English?

The language and style of characteristic of Wikipedia articles is determined a variety of factors. From the perspective of linguistic variability, however, two issues seem particularly relevant: the linguistic background of the editors, and the usage notes in the *Manual of Style*. Based on both these factors, we could expect American English variants to be dominant in the English Wikipedia across the board. First of all, according to April 2011 *Editor Survey*, 20% of Wikipedia editors reside in the United States, as opposed to a 6% who are based in the UK (Pande 2011: 39). On the other hand, Wikipedia is generally neutral with regard to national varieties of English. The style notes promote consistency within articles and recommends the use of terms and spellings that are common to all varieties of English. In addition, it discourages disputes concerning the appropriate variety to be used in articles. [10]

At the same time, Wikipedia articles do not form a uniform collection, and we can expect to find linguistic variation between articles of different kind. [11] On the one hand, there is an enormous variety between the kind of editors who contribute to Wikipedia, representing a variety of backgrounds, interests and levels of commitment. The *Editor Survey 2011* observes that Wikipedia is typically edited by computer-savvy college graduates in their thirties who live in Europe or in the US, but that

> ‘typical’ doesn’t tell the whole story. Our community comes from a widely varied set of backgrounds, and requires thoughtful and sensitive interactions within the community – because the person behind the username is quite likely different from you. (Pande 2011: 2)

Like editors, individual articles can also be very different from each other. In general, different types of articles tend to attract different types of contributors, and it is often easy to predict how this works. For example, Myers (2010: 139–40) observes that many (though not all) editors of the article [[Manchester]] come from Yorkshire or have an interest in some aspect of the city. But there are clearly other factors at work, too: Myers further notes that articles may also follow radically different paths of development and attract different types of contributors, depending on whether they are perceived as controversial.

The *Manual of style* also identifies a few situations where the choice of national variety actually makes a difference. These include verbatim quotations, proper names (persons, titles of books etc.), comparisons of varieties of English, and most importantly, articles with ‘strong ties to a particular English-speaking nation’, which should employ the variety associated with that nation.

Against this background, we can formulate hypotheses about the prominence of national varieties in the English Wikipedia. On the one hand, the large number of US-based Wikipedia editors would suggest that American English usage is dominant in the articles, as previously described. However, the choice of the variety of English may also depend on the topic of the articles. In particular, if the editors follow the guidelines set forth in the *Manual of Style*, we can also expect British English to be used in articles on the geography, institutions and culture of the United Kingdom. Using a corpus-based approach, it is possible to test these hypotheses against authentic data. To my knowledge, the use of national varieties in Wikipedia articles has not been systematically investigated from a variationist perspective. [12]

4. Corpus

Section 2 listed several corpora compiled from Wikipedia articles. In this study, I use the *Westbury Lab Wikipedia Corpus (2010)* (Shaoul and Westbury 2010). [13] It is the largest of these corpora and can therefore provide a wealth of data about the choice between different variants. However, using the *WLWC* is also less straightforward than most other published corpora, and therefore it needs to be processed before using it to address the research questions stated introduced in the previous section. [14]

The approach followed in this article is to extract different samples from the main corpus and compare them with respect to the prominence of different spellings and grammatical structures. As the WLWC is released as a plain text corpus, it provides no information about the hierarchical relationships and the links between the articles. For this reason, the extraction of the samples was based on the titles of the articles.

The *WLWC* is provided as one large text file – over 6 GB – without ready-made internal divisions according to parameters as text type or topic, familiar from such corpora as the *British National Corpus* and the *Corpus of Contemporary American English*. Such a file is far too large to be opened in a standard concordance program such as AntConc, or a in a text editor. [15] For this reason, the corpus first needs to be converted into a format that is more easily manageable.

The only annotations in the corpus are the article boundaries. [16] Using these boundaries, an R script was written to split the massive file into smaller files, each of which contains one article. The total number of such articles is 29,990,323. A list of these articles was saved in a separate text file, together with the word count of each article. This list can be used as an index to the corpus, and with the help of it, it is possible to extract subsets based on article titles and their word counts.

4.1 Article length in the *WLWC*

The range of article length is considerable in the *WLWC*, from 2 to 52,500 words (the longest article is *USS Wallace L. Lind (DD-703)*). [17] This can be observed in Table 1, which prints the deciles of the lengths of articles.

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
2	23	39	60	91	133	192	280	422	741	52,500

Table 1. Deciles of article length in the *WLWC* (measured as word tokens).

Information about article length is useful and needs to be taken into account when choosing representative samples from the corpus. As shown in Table 1, the majority of articles in the *WLWC* are fairly short: alongside a relatively small number extremely long articles, ninety per cent of all texts in the corpus in fact only contain 741 words or fewer. The mean length of article is 323 words. In addition, many articles are extremely short, comprising only one sentence. [18] The most typical value of article length is 12 word tokens, found in over 33,000 articles.

Very short texts are often excluded from corpora. Ide et al. (2002), for example, only accepted texts containing more than 2,000 tokens to form part of the *American National Corpus*, because only such texts would be suitable for such tasks as collocate

extraction, concordancing, and the analysis of syntax and discourse.

There are good reasons for excluding very short articles from this study. While typical in the sense that they are extremely common, Wikipedia editors or readers would hardly regard them as good examples of articles. Some of them are mere disambiguation pages providing links to other articles, illustrated in Example 1, while others represent a starting point of article which is ideally expanded in future (Example 2).

(1) **Hostel (disambiguation)**.
Hostel is a budget-oriented overnight lodging place. (WLWC: *Hostel*)

(2) **Hermershausen**.
Hermershausen is a Stadtteil of Marburg in Hesse. (WLWC: *Hermershausen*)

Moreover, as the *WLWC* also contains a considerable number of long texts, it is safe to exclude texts like (1) and (2) in order to obtain a sample that is more useful for corpus linguistic analysis. This study only considers articles with a word count of 1,000 words or more. Even if this means excluding over 90 per cent of the articles in the Westbury Lab corpus, the remaining part of the corpus still contains over 196,000 texts and 412 million words, a large enough sample for analysing both syntactic and lexical questions.

4.2 Extraction of samples

To investigate the question of variety, three samples were extracted from the *WLWC* (excluding texts shorter than 1,000 words). Sample 1 is a random sample of 1,000 articles. Sample 2 represents articles containing ties with the United Kingdom. A list of such articles is obtained through searching for article titles containing any of the following items: *United Kingdom*, *the UK*, *Great Britain*, *British*, *England*, *Briton*, *the BBC*, as well as all town names included in the Wikipedia articles [[List of Towns in England]] and [[List of Cities in the United Kingdom]]. [19] As these criteria apply to several thousand articles, Sample 1 contains 1,000 randomly selected articles from this list. Sample 3 represents articles with ties with the United States, and is randomly extracted from all the articles which contain the string *United States* in the title, as well as all the names of the cities with a population over 100,000 listed in the article [[List of United States cities by population]]. The details of the three samples are summarised in Table 2.

Sample	Type	Texts	Words
Sample 1	Random sample	1,000	2,345,582
Sample 2	Articles with ties to the UK	1,000	2,912,891
Sample 3	Articles with ties to the US	1,000	3,434,385

Table 2. The samples used in the analysis

5. Method of analysis

This article investigates variability on two levels of linguistic analysis, spelling and grammar. It should be noted that as the data represents written language, it is not possible to analyse the most obvious differences between UK and US English – intonation, stress patterns, and the articulation and distribution of vowels (Algeo 2006: 2) – which can only be observed in spoken language. While spelling and grammar differences are much less dramatic in comparison (see Mair 2009a), it has recently been claimed that there are in fact many more differences between the grammars of these two main varieties than are commonly recognised (Rohdenburg and Schlüter 2009a: 1–2). As Rohdenburg and Schlüter put it, 'contrary to general opinion, BrE and AmE do not differ only in their pronunciation and lexicon, but also in central domains of their grammar' (2009b: 364).

As previously mentioned, this study employs a variationist approach. In other words, it is concerned with what the different ways of saying (roughly) the same thing are, and what their probabilities of occurrence are in different contexts (e.g. Jucker 1992: 18–20; Barbieri 2005: 224; Tagliamonte 2006: 10–14). Accordingly, each of the variables analysed below – spelling, concord with collective nouns, preposition choice after *different* and preterite forms of some verbs – are defined in such a way that that they comprise all the paradigmatic choices available at that particular level. The aim of this study, then, is to compare how the realisation of each dependent variable is predicted by the independent variable, whose values correspond to ties that the article has to different English-speaking countries (see Section 4.2). Such a comparison provides data to investigate the hypothesis formulated in Section 3 that articles in general prefer American English variants, while variants associated with British English would be dominant in articles with ties to the UK.

Each variable is analysed individually in this study, and all comparisons are carried out at the level of subcorpora. Accordingly, the relative frequencies of variants in each sample is counted by combining the observations from all texts. The Chi-square test is

used to assess the null hypothesis (H_0) that there is no difference between the rate at which American and British English variants are used in the three samples, with the significance level set at α=.05.

5.1 Spelling

To investigate whether the samples follow British English or American English spellings, each sample was searched for words known to have distinct spellings in these varieties. A number of such lists have been created for different purposes and made available on the internet. [20] In this study, I use *VarCon* (*Variant Conversion Info*), a table listing British, American and Canadian spellings of words, created by Kevin Atkinson (Atkinson 2011). As well as with being recently updated, *VarCon* is also better documented and more flexible than most other similar lists. [21] For instance, it contains information about how common each word is, which would enable filtering out very infrequent spellings, if considered necessary.

In total, *VarCon* lists over 18,000 forms whose spelling is different in at least one of the varieties in question. The large number of items on this list is explained by the fact that it includes complete lemmas of low-frequency verbs, such as *Prussianise*/*Prussianise* and *Platonise*/*Platonize*, as well as nominalisations derived from them. For this study, only those items were used whose preferred BrE spelling and the preferred AmE spelling were indicated to be different. This subset contains 15,435 items with variant spellings.

Obviously, the vast majority of these words are extremely unlikely to occur in the corpus data. This can be seen in Table 3, which provides a random sample of 20 words on the list, containing such words as *tantalisingness*, *platitudiniser* and *sanctuarises*. [22] However, the analysis only takes into account those instances which are actually found in corpus data, and the non-occurrence of items does not therefore have an impact on the results. For this reason, low-frequency items were not removed from the list at this stage.

BrE	AmE
harmonisablest	harmonizablest
Sovietising	Sovietizing
solemnising	solemnizing
catholicising	catholicizing
Lutheranise	Lutheranize
palaeoethnography	paleoethnography
phenolisation	phenolization
terroriser’s	terrorizer’s
parenthesises	parenthesizes
demoralisation	demoralization
heroisation’s	heroization’s
tantalisingness	tantalizingness
lithaemia	lithemia
haematogenesis	hematogenesis
decolonising	decolonizing
platitudiniser	platitudinizer
trivialisations	trivializations
supersensitisation	supersensitization
mahoganised	mahoganized
sanctuarises	sanctuarizes

Table 3. Sample of words with variant spellings based on *VarCon* (for more information, see Atkinson 2011).

While the list appears to be a rather comprehensive representation of spelling differences between these two varieties of English, at least as far as its size is concerned, it is actually a rather crude simplification of a complex situation, and given that it was created semi-automatically from an on-line source, it is not intended as a definite and or even error-free inventory of variant spellings. Moreover, as the corpus consists of unannotated plain text, it is not possible to control for such things as whether the word occurs in a direct quotation, or as part of a proper name whose spelling is fixed (e.g. *Labour Party*). However, such inaccuracies are likely to be offset by the large total number of items on the list, and the frequencies are certainly indicative of the preference for one national standard over another, despite the limitations of the list. Therefore, we treat the *VarCon* subset as a sufficiently accurate

representation of words with variant spellings in the main varieties of English.

5.2 Complementation

In addition to spelling, I shall investigate some patterns of grammar and complementation which are known to vary across the varieties of English. These include:

Agreement with collective nouns. Singular collective nouns may be followed by either singular or plural verb forms. While most nouns take singular concord, plural concord is common in British English with some nouns, including *family*, and *crew*, *army*, *committee*, and *government* (see Biber et al. 1999: 188, Hundt 1998: 80–88, Hundt 2009: 27–30, and Algeo 2006: 279–286). Nouns referring to sport teams (e.g. *Leeds* – [[Leeds United A.F.C]] or *Liverpool* – [[*Liverpool F.C*]]) regularly take plural concord in British English (Algeo 2006: 280). In this study, six such words are analysed with respect to concord.

Choice of preposition after *different* The adjective *different* may take three prepositions in Present-day English: *from*, *than*, and *to*. Based on previous studies, *from* is the main variant, occurring most frequently in both British and American English. The alternatives *than* and *to* are distributed differently, the former being more frequent in AmE, and the latter almost exclusively used in BrE (see e.g. Hundt 1998: 107, Algeo 2006: 258, and Mair 2009a: 89).

Regular/irregular inflected forms of some verbs. Some verbs have both regular and irregular variants of preterite and past participle forms: *dreamt*/*dreamed*, *leapt*/*leaped*, *spelt*/*spelled*, etc. According to Biber et al. (1999: 396–397), the relative preference of these variants depends on register and grammatical use, but in general American English shows a stronger preference for the regular variant of many of these verbs compared to British English [23] (see also see also Algeo 2006: 16–19 and Levin 2009). [24] In this study, I shall investigate the preterite and past participle of the same verbs that were analysed in Levin (2009): *burn*, *dream*, *dwell*, *kneel*, *lean*, *leap*, *learn*, *smell*, *spell*, *spill*, and *spoil*.

6. Results

6.1 Spelling

Table 4 shows how the BrE and AmE variants of the words on the search list are distributed across the three samples.

Sample				
Variant	**Sample 1**	**Sample 2**	**Sample 3**	**Total**
BrE	4,225	16,196	1,489	21,910
AmE	8,728	2,250	22,466	33,444
Total	12,953	18,446	23,955	55,354

Table 4. Spelling of words in three samples from the *WLWC*.

To compare the tendencies across the samples, it is convenient to visualise this data as a mosaic plot, provided as Figure 1. Each tile in the plot represents one cell in Table 4. The size of a tile is determined by the count of the corresponding cell in relation to the other cells. The colour of the tile represents the Pearson residual of the cell, [25] which helps us identify individual cells with a significantly higher than expected observed frequency – highlighted in blue – and those with a significantly lower observed frequency – highlighted in red. [26]

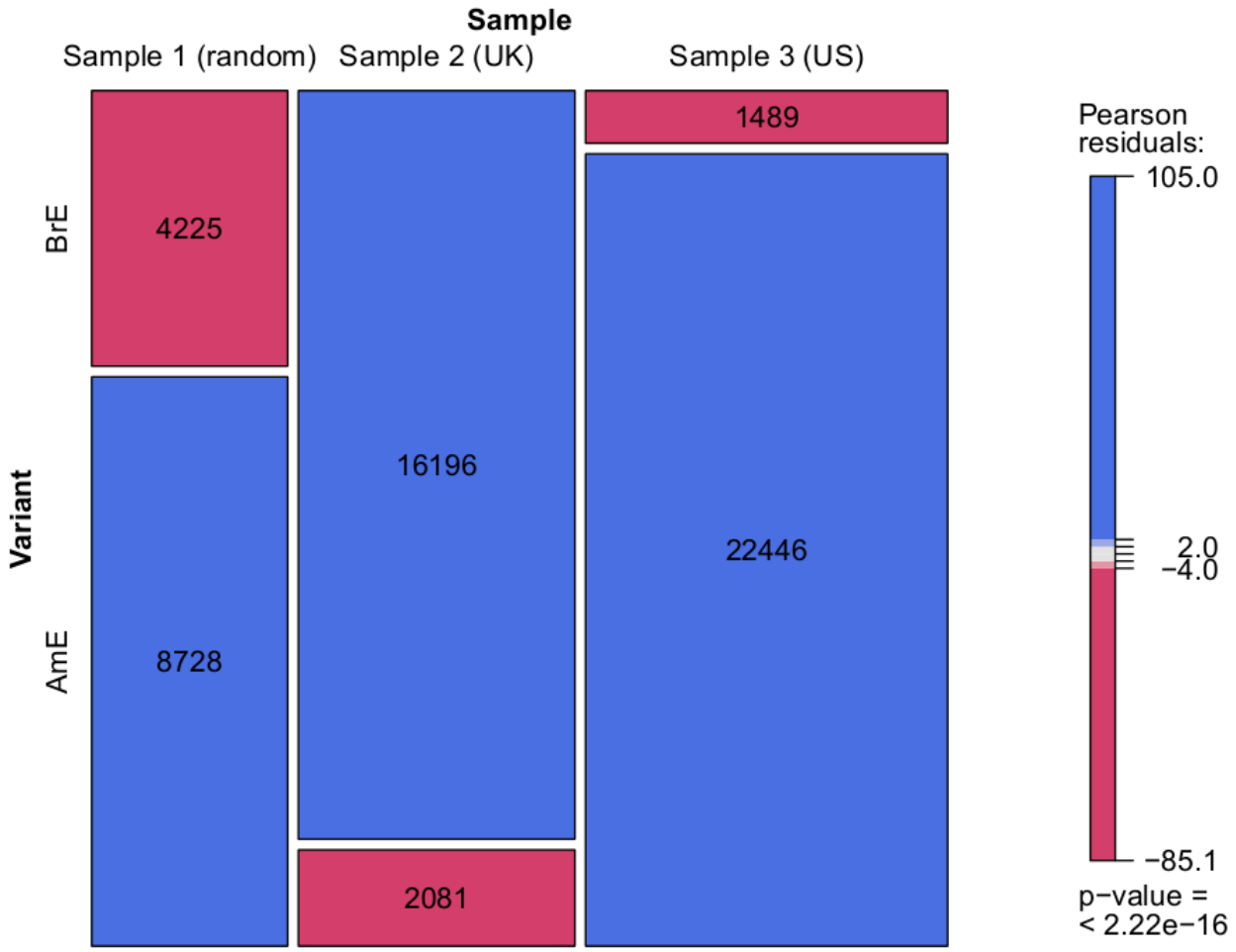

Figure 1. Spelling variants in three samples from the *WLWC*.

Figure 1 confirms the hypothesis that Sample 2 favours British English and Sample 3

American English spellings. The difference is extremely clear: the BrE spelling is used in 87% of the occurrences in Sample 2, but only in 6% in Sample 3. The randomly chosen Sample 1 is situated between these two extremes, but shows a clear preference for AmE spellings (67%). The cell recording the number of BrE spellings in Sample 2 shows the largest deviation from the expected value (16,703 vs. 7,231), but as the shading in the figure shows, each cell individually violates the null hypothesis of independence. The distribution is highly significant (χ^2=29349.40, df=2, p<0.001).

To illustrate the difference in style between Sample 2 and Sample 3, an extract is quoted from each. Example 3, taken from the article *British Army during the Second World War*, employs the BrE spellings *-ise* and *-our* consistently. By contrast, the extract from a military-themed article from Sample 3 (Example 4) shows the opposite tendency, favouring *-ize* spellings.

(3) In late 1940, following the campaign in France, the divisions were **reorganised** on paper. It had been **realised** that mixing light and cruiser tanks in the same brigade had been a mistake and that there were insufficient infantry and support units within the division. Each **armoured** brigade now incorporated a **motorised** infantry battalion, and a third battalion was present within the Support Group. (Sample 2: *British Army during the Second World War*)

(4) The task of **organizing** the U.S. Army commenced in 1775. During World War I, the "National Army" was **organized** to fight the conflict. It was **demobilized** at the end of World War I, and was replaced by the Regular Army, the **Organized** Reserve Corps, and the State Militias. (Sample 3: *United States army*)

The hypothesis that the spelling reflects the subject matter of the article is clearly borne out by the data. This finding also confirms that the chosen approach is in general useful for determining the extent to which articles conform with a specific national standard.

6.2 Concord with collective nouns

To investigate whether collective nouns take singular or plural concord in the data, six verbs were selected for analysis: *army*, *team*, *family*, *public*, *government*, and *committee*.

A search for these wordforms in the three samples retrieves a large number of

occurrences, but most of them do *not* in fact involve a choice between singular/plural concord and therefore need to be discarded. Due to the poor precision of this type of lexical searches in a plain text corpus, the present analysis only takes into account those instances where the collective noun is the head of a noun phrase and it is immediately followed by a verb form, which given the context could either be singular or plural. It is necessary to introduce this criterion to keep the amount of manual work manageable, in spite of the fact that it is likely to somewhat reduce the recall.

The distribution of singular and plural verb forms found in these specific contexts is shown in Table 5 and represented graphically as a mosaic plot in Figure 2.

	Sample			
Concord	**Sample 1**	**Sample 2**	**Sample 3**	**Total**
Singular	242	461	561	1,264
Plural	22	59	18	99
Total	264	520	579	1,363

Table 5. Singular and plural concord with six collective nouns in three samples of the *WLWC*.

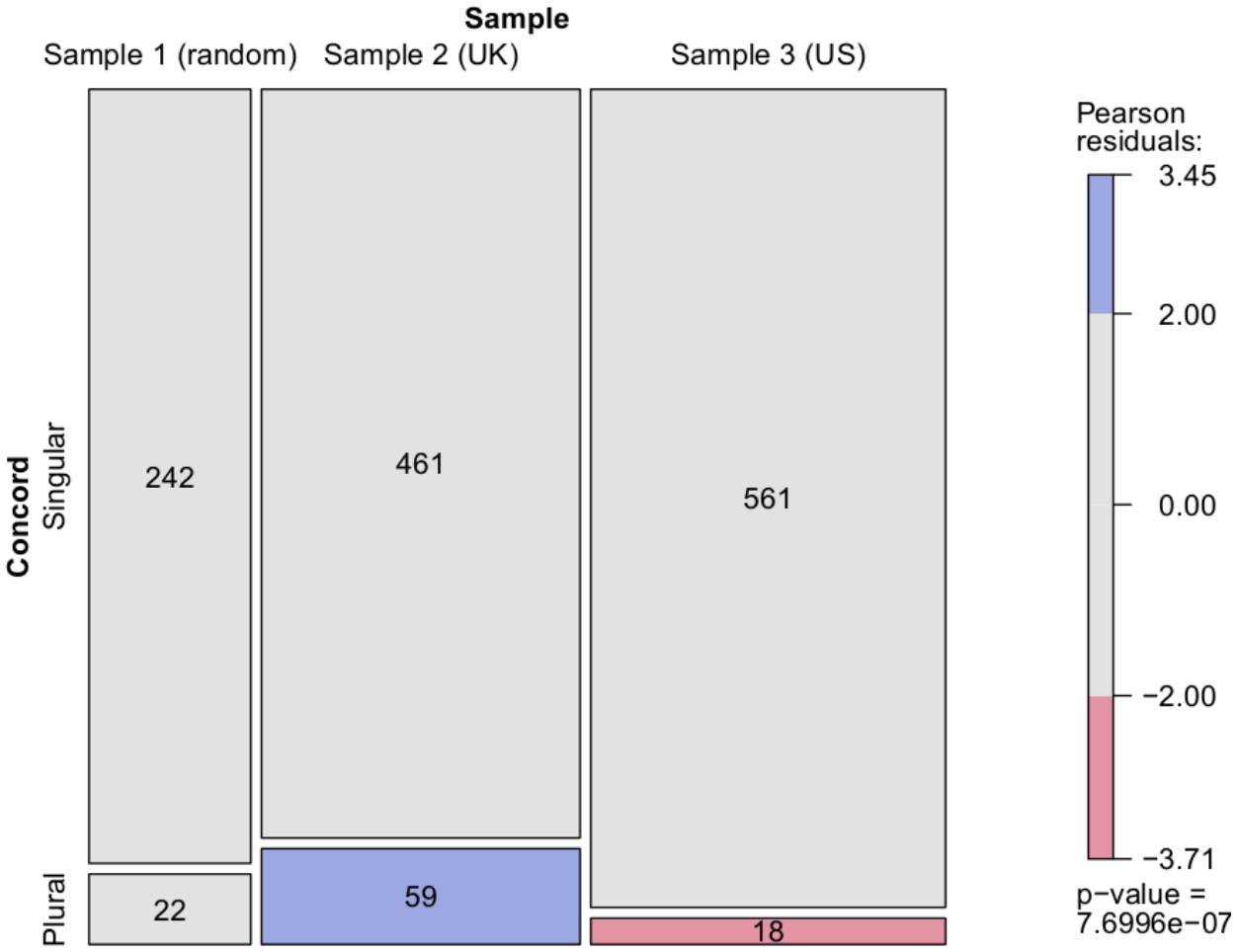

Figure 2. Singular and plural concord with six collective nouns in three samples of the *WLWC*.

As can be observed, singular concord is the norm in all three samples. However, plural concord, which is said to be characteristic of British English, is proportionally most common in Sample 2, which represents articles with ties to the UK. The shading in Figure 2 indicates that the observed frequency of the relevant cell is significantly higher than expected (59 vs. 38 instances), and that the opposite is true for Sample 3. Examples (5) and (6) illustrate the use of plural concord in Sample 2:

(5) Ipswich's sole professional football team **are** Ipswich Town, …(Sample 2: *Ipswich*)

(6) The British government **were** determined that British territories, such as Hong Kong, should be recaptured by British forces. (Sample 2: *British Pacific Fleet*)

The difference between the three samples is statistically significant (χ^2=28.15, df=2, p<0.001). However, compared to the results of spelling differences (Section 6.1), the effect is obviously much weaker.

Table 6 shows the distribution of the prepositions *from*, *than* and *to* in Samples 1–3, and the same data is presented as a mosaic plot in Figure 3.

	Sample			
Variant	**Sample 1**	**Sample 2**	**Sample 3**	**Total**
from	49	61	48	158
than	16	1	16	33
to	3	15	1	19
Total	68	77	65	210

Table 6. Prepositions following *different* in three samples from the *WLWC*.

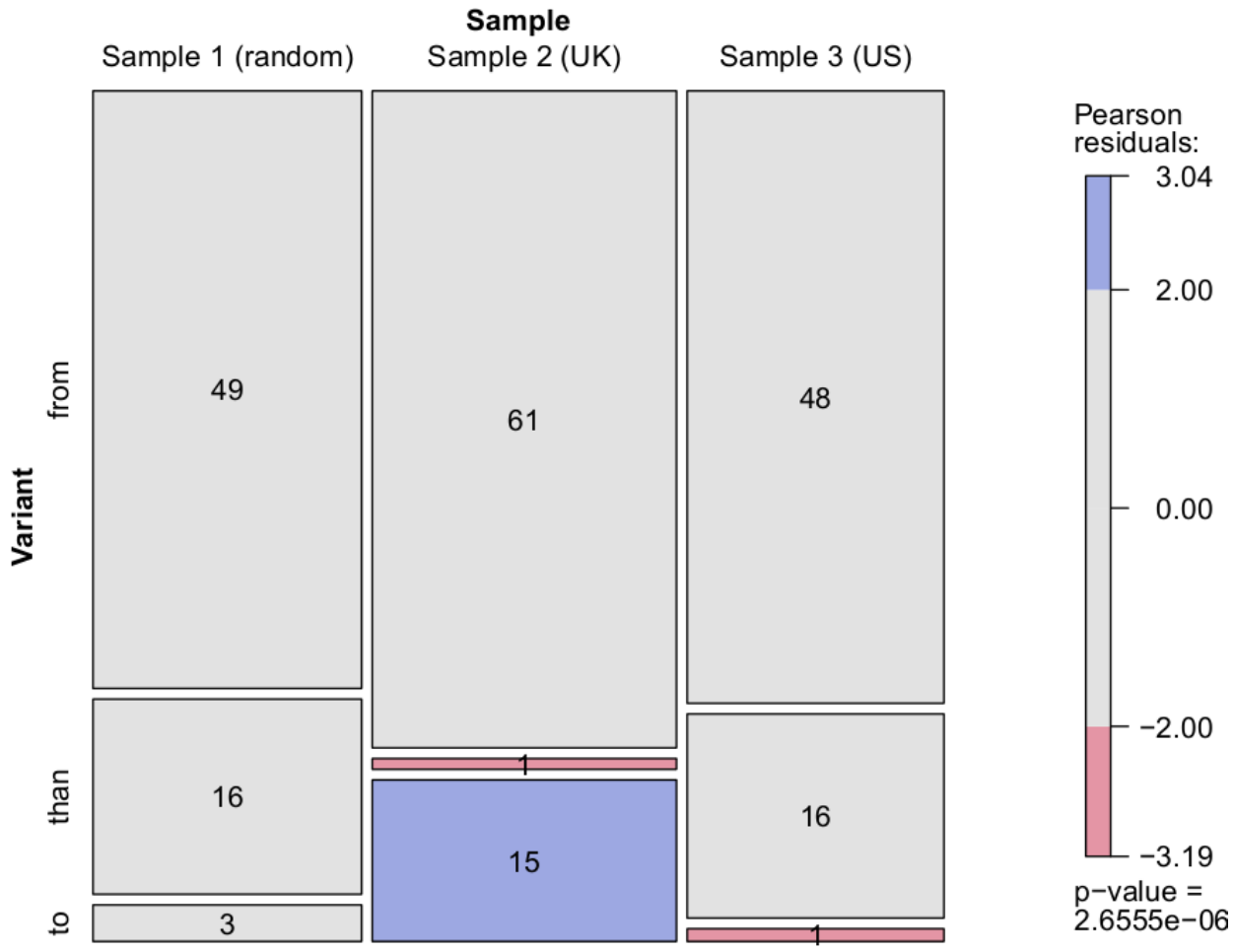

Figure 3. Prepositions following *different* in three samples from the *WLWC*.

As can be seen, the data conforms to our expectations remarkably well. First, *from* is clearly the preferred preposition across the board, accounting for well over two-thirds

of the instances in all three samples. In addition, the secondary choices *than* and *to* are distributed as expected. The relative frequencies of the variants are almost identical in Samples 1 and 3, where *than* is chosen roughly once in every five occasions. The use of *to*, meanwhile, is virtually limited to Sample 2, where it covers one fifth of the total number of occurrences, significantly more than expected (see the shading in Figure 3). This usage is illustrated in Example (7). The variant *than* is correspondingly less frequent than expected in this Sample 2, and more frequent than expected in Sample 3. Example (8) illustrates this usage. The difference of distributions in Samples 1–3 is statistically significant (χ^2=31.3028, df=4, p<0.001).

(7) All continental samples were statistically different **to** British samples. (Sample 2: *Anglo-Saxon settlement of Britain*)

(8) Berlin was an occupied city, with a status very different **than** any other part of East or West Germany. (Sample 3: *Embassy of the United States in Berlin*)

Despite the moderately low frequency of the secondary prepositions in particular, it can be observed that the proportions in Table 5 are similar to those reported in previous studies, including Hundt (1998: 107), based on newspaper data, and Mair (2009a: 109), based on counts of Google hits from different Anglophone domains. This data provides corroborating evidence that Samples 2 and 3, which were selected based on their subject matter, employ British and American English, respectively.

6.4 Regular/irregular preterites and past participles

Table 7 shows the distribution of the regular preterite/past participle forms of nine verbs – *burn*, *dwell*, *lean*, *leap*, *learn*, *smell*, *spell*, *spill*, and *spoil* – across the three samples. [27]

Sample							
	Sample 1		Sample 2		Sample 3		
Verb	-ed	-t	-ed	-t	-ed	-t	Total
burn	52	22	38	37	54	4	207
dream	4	2	0	0	0	0	6
dwell	0	0	0	0	1	0	1
kneel	0	0	0	0	0	0	0
lean	5	0	1	0	6	0	12
leap	4	4	1	0	1	3	13
learn	87	21	36	10	70	1	225
smell	0	0	1	0	2	0	3
spell	7	0	5	0	9	0	21
spill	3	0	4	0	6	0	13
spoil	4	0	2	0	5	0	11
Total	166	49	88	47	154	8	512

Table 7. The use of regular and irregular verb forms in three samples from the *WLWC*.

As can be seen, the regular *-ed* forms are more frequent than the irregular *-t* forms in all samples, based on the total number of occurrences. However, *-t* forms are proportionally far more common in Sample 2 than Sample 3, which conforms with our expectations, based on the recommendation that articles with ties to the UK should employ British English.

It should be noted, however, that genre and the subject matter of the texts seem to offer few opportunities for using most verbs listed in the table. The row totals are dominated by two verbs, *burn* and *learn*, for which the distribution of variants is shown as a mosaic plot in Figure 4. In fact, these are the only verbs providing meaningful data for statistical analysis, as the other verbs either have an extremely low frequency (<10 for *dream*, *dwell*, *kneel*, and *smell*), and in the case of *spell*, *spill*, *spoil*, and *lean*, show no variation between *-ed* and *-t* forms in the samples. [28] That said, Figure 4 clearly demonstrates that *burnt* and *learnt* are significantly more frequent in Sample 2 than Sample 3, which is in accordance with the original research hypothesis.

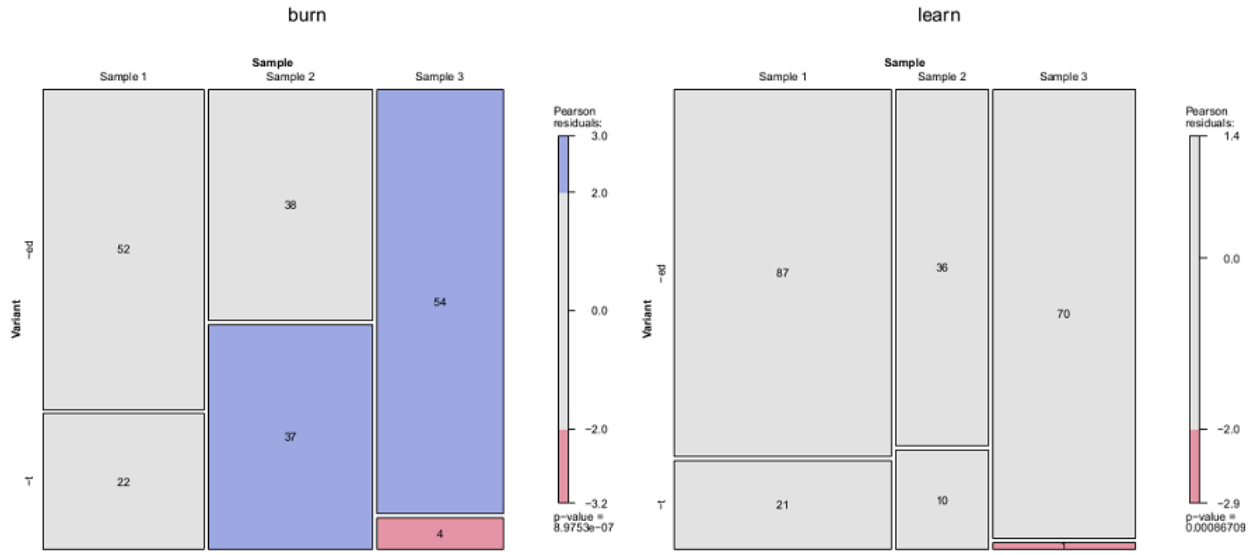

Figure 4. Distribution of variants *-ed* and *-t* for *burn* (left panel) and *learn* (right panel) in three samples from the *WLWC*.

There is one anomaly in that data presented in Table 6: the incidence of *learnt* is roughly the same in Samples 1 and 2. This is somewhat surprising, as we would have expected the BrE form *learnt* to have a higher incidence in Sample 2, which contains articles with ties to the UK. A closer look at the data shows that that Sample 1 contains one outlier which dominates the result: the article on *Aala Hazrat* (Islamic scholar) contributes ten instances of *learnt* to the aggregate count. Example (9) is quoted from a short section within this article, *Early life and Education*, which contains in total five instances.

(9) He **learnt** 21 subjects from his respected father, Naqi Ali Khan (d. 1880). (Sample 1: *Aala Hazrat*)

Incidentally, this article has subsequently been renamed as [[Ahmed Raza Khan Barelvi]], and the section in question is no longer found in it (accessed in June 2014) – it was removed on 12 May 2010, roughly a month after the article had been included in the *WLWC*. [29] The current version contains a much shorter section entitled 'Early life', and the sentence about the influence of the scholar's father now provides more details and adopts a style that is more appropriate to an encyclopedia article (Example 10):

(10) He studied Islamic sciences and completed a traditional Dars-i-Nizami course under the supervision of his father Naqi Áli Khan, who was a legal scholar. ([[Ahmed Raza Khan Barelvi: Early life]])

On a more general level, these observations highlight one problem associated with the

approach followed in this paper, *ie* simple pooling of the data. While the chosen approach is well suited for the analysis of evenly distributed high-frequency phenomena like spelling variation (Section 6.1), it is clearly less reliable with low-frequency items. In such circumstances, other approaches would likely provide more reliable estimates of the relative frequency of the variants (see e.g. Hinneburg et al. 2007).

7. Discussion

The results presented in Sections 6.1–6.3 provide strong evidence for the idea that Wikipedia articles in general favour American English usage norms. Sample 1, which consists of a random sample of 1,000 articles (over 2.3 million words), prefers AmE spellings and regular preterite forms, and employs singular concord with the chosen collective nouns, and *than* as the secondary choice of preposition after the adjective *different*. These tendencies are similar to those observed for Sample 3 (articles with ties to the US), although less extreme, as can be seen.

By contrast, both of these samples are markedly different from Sample 2, which represents articles with ties to the UK, with respect to all four variables analysed in this article. Accordingly, Sample 2 prefers BrE spellings and shows a higher incidence of irregular preterites, plural concord with collective nouns, and the use of *to* after *different*. This finding suggests that UK English seems to hold its own also in Wikipedia (cf. Atwell et al. 2007: 5).

It is likely that the reason why AmE is dominant in Wikipedia is linked with the demographic characteristics of editors: US-based editors, who can be expected to follow American English usage norms in their contributions, clearly outnumber the editors from other Anglophone countries (Pande 2011: 39), and the findings for Samples 2 and 3 seem to reflect this situation. At the same time, the results also suggest that the *Manual of Style* has an influence on the articles. While the *Manual* primarily recommends using common-core variants when possible, it also allows the use of regionally marked variants in articles dealing with specific nations and their culture. Given that Samples 2 and 3 are very different with respect to their preferred variants, it therefore seems that many editors are indeed familiar with the *Manual*'s recommendations and use them to edit articles and resolve possible disputes about style.

While the observed tendencies are clear, we should nonetheless remember they are based on the analysis of corpora alone, and as such do not directly tell us why a certain variant is chosen in a particular situation. Of course, this caveat applies to the use of

corpus data in general: using Widdowson's (2000: 6) terminology, corpora are 'third person observed data', describing usage from the analyst's perspective. The limitation of this type of data, according to Widdowson, is that

> it cannot represent the reality of first person awareness. We get third person facts of what people do, but not the facts of what people know, nor what they think they do: they come from the perspective of the observer looking on, not the introspective of the insider… [C]orpus analysis deals with the textually attested, but not with the encoded possible, nor the contextually appropriate. (2000: 6–7)

For this reason, it would be extremely useful to complement the analysis with data obtained using other methods. For example, while no attempt has been made in this article to investigate what individual articles are like and who has contributed to them, future studies could look at individual articles in more detail and consult their version histories to find information about their development. [30]. In addition, the investigation of specific editors' linguistic choices using multivariate methods might provide further insigts into the factors determining the choice. It would also be interesting combine the analysis of corpus data with the information about the interaction between editors, as manifested in the 'Talk' pages (cf. Pentzold and Seidenglanz 2006; Myers 2010). Future studies could explore ways of doing this in a systematic fashion.

By the same token, the analysis of corpus data can also be improved on in future work. For one, while the pooling of observations from different texts — a commonly adopted approach, and the one used in this study — provides information about the overall differences between samples, it also makes the assumption that they are evenly distributed between the texts, which is not necessarily true in all situations (see Section 6.4). It could be possible to refine the results by considering other approaches of sampling the data and estimating the frequencies, which would also take into account the dispersion of the observed features in the samples (e.g. Hinneburg et al. 2007; Gries 2006; Gries 2008). Identifying and annotating named entities prior to the analysis of data would also improve the accuracy of the findings.

Methodologically, the present study demonstrates the usefulness of the variationist framework in the analysis of Wikipedia data. Specifically, the fact articles on the United Kingdom (Sample 2) and the United States (Sample 3) both employ regionally marked forms in predictable ways indicates that the chosen approach can be applied to any group of Wikipedia articles to find out about its preference for a particular national variety. To illustrate this point, Figure 4 shows the preferred spellings in three small samples of Wikipedia articles, which contain articles associated with three

Nordic countries, Finland (FIN), Norway (NOR), and Sweden (SWE). [31] As can be seen, the distributions are very similar in all three samples, showing a slight but consistent preference for American spellings (56%–58% of the cases). [32] It is interesting, however, that the incidence of BrE variants in each of these three samples is higher than in the randomly selected Sample 1 (see Figure 5), which suggests that they would in fact prefer BrE forms more than Wikipedia articles in general.

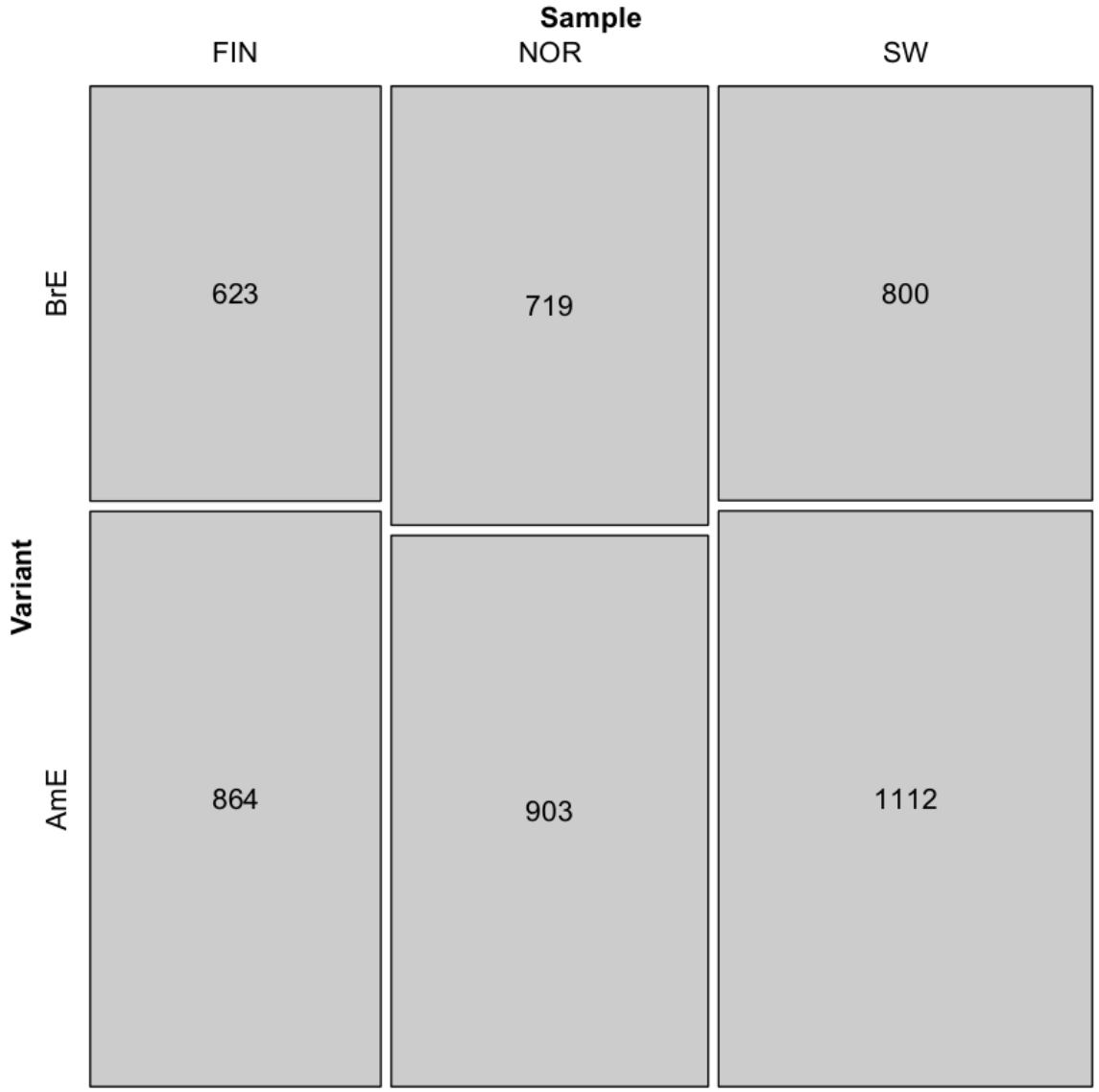

Figure 5. Spelling variants in Wikipedia articles with ties to Finland, Norway and Sweden.

This finding opens many new questions, including how consistent this preference is, and what gives rise to it. If we further assume the that these articles are edited by advanced learners of English, who are native speakers of the main languages in these countries, we could look into what varieties are preferred in these countries in general. In Sweden, for example, Standard British English has been the preferred model in EFL instruction throughout the second half of the 20th century (e.g. Modiano 1993), and it seems that this preference has not changed much in recent years, despite claims of

increasing Americanisation (Larsson 2012). A more thorough analysis of these issues, too, is left for further research.

Finally, the findings of this study also have some implications for the use of Wikipedia as a teaching tool. Previous studies have alrealy shown that producing and editing Wikipedia articles in the context of academic writing instruction is a useful way of teaching academic literacy and writing skills (Tardy 2010, Miller 2012). Wikipedia articles are created through a process in which meanings are negotiated collaboratively, which bears many similarities to how academic papers are written and revised. The present results show that Wikipedia articles need to conform with the recommendations in the *Manual of Style*, in order to be successful. Therefore, practical activities drawing attention to specific points in the *Manual* – e.g. adding new content to Wikipedia, or editing existing articles so that they would conform with the recommendations – are likely to help students when they revise their research reports and term papers, which are stylistically similar to Wikipedia articles (Konieczny 2007).

8. Conclusion

This study has investigated the consistency of preferences for a particular national variety across three samples of 1,000 Wikipedia articles. As Myers (2010: 23) observes, each successful Wikipedia article has been produced by a community of people who shares some basic norms about the content, form and and style of the article. Against this background, the main findings of this study – that Wikipedia articles about the US tend to follow AmE usage norms, while BrE is preferred in articles about the UK – are perhaps not all that surprising. From another perspective, however, the fact the samples show such a consistent preference for one national variety over the other is in fact remarkable, taking into account that this is not necessarily true for all web texts. For example, Atwell et al. (2007) report that while on-line texts downloaded from Australian, South African or Irish domains tend to be closer to UK than to US English, it is often difficult to see a clear preference for a particular variety over another (see also Atwell et al. 2009).

The findings of this study therefore underline the fact that compared to web texts in general, Wikipedia articles form a rather homogeneous text collection. In earlier studies, it has been suggested that this homogeneity is in part created by the fact that the editors follow the neutral point of view principle, one of Wikipedia's three core content policies, which requires that all significant views a given topic should be represented fairly, proportionately, and without bias (Emigh and Herring 2005). The analysis presented in this article suggests that the *Manual of Style*, one of Wikipedia guideline documents, operates in a similar way. By providing recommendations on

matters of style, which the editors clearly take very seriously, the *Manual* also contributes to creating linguistically and stylistically homogenous articles.

Notes

[1] Wikipedia articles referred to in this article are enclosed within double square [[ ]] brackets.

[2] See http://stats.wikimedia.org/wikimedia/squids /SquidReportPageViewsPerCountryOverview.htm.

[3] A wiki is a web page that can be edited by anyone who accesses it, see e.g. ([[Wiki]])

[4] See Myers (2010: 129–130, 143–144) for a useful overview of these criticisms.

[5] Jimmy Wales shares this opinion (Young 2006).

[6] However, Kolbitsch and Maurer (2006: 196) point out that this policy can be interpreted differently depending on writers' cultural, social, national and linguistic backgrounds. Cultural differences have also been found between different Wikipedia editions; Callahan and Herring (2011), for example, found that compared to the Polish edition, the English Wikipedia follows traditional encyclopedic norms more closely.

[7] The *Wikipedia XML* corpus has texts in eight languages: English, French, German, Dutch, Spanish, Chinese, Arabian and Japanese, and the *Wikicorpus* in three: Catalan, English and Spanish.

[8] This number was calculated before the text was converted to plain text (Cyrus Shaoul, personal communication). The corpus therefore includes a large number of very short articles (see Section 4.1 and Table 1).

[9] WaCkypedia_EN is available on http://wacky.sslmit.unibo.it /doku.php?id=corpora

[10] See http://en.wikipedia.org /wiki/Wikipedia:Manual_of_Style#National_varieties_of_English.

[11] In fact, some commentators have paid attention to the variability of English in Wikipedia: Dalby (2007), for instance, observed in 2007 that while the articles in general adopted an encyclopedic style, many articles mixed American and British English norms instead of consistently following one or the other. However, as Dalby provides no examples of such inconsistent articles, it is not possible to determine whether their style has subsequently been edited to better conform with the norms of

either variety.

[12] The observations in Dalby (2007: 7), mentioned in note 11, appear to be based on the author's impressions rather than a systematic inquiry.

[13] Henceforth *WLWC*.

[14] However, the WLWC contains no mark-up tags or boilerplate, and is therefore much easier to use than web texts that have not been 'cleaned up', see e.g. Kehoe and Gee (2007).

[15] Even if the file can be viewed using Large Text File Viewer, it is not convenient to use it for searching or editing.

[16] Each article ends with the string `--END.OF.DOCUMENT--`.

[17] An article whose length is two words contains no text, only the title and the code `_NOTOC_`, standing for "no table of contents".

[18] See also Section 2 and footnote 8.

[19] Articles which obviously refer to non-UK places and people, e.g. New England and the British Columbia, were removed from the list.

[20] See, for example, the *Word list of UK and US spelling variants by Words Worldwide Limited* (http://www.wordsworldwide.co.uk/articles.php?id=37) and Roland Grant's *Comprehensive list of American and British spelling differences* (available on http://www.tysto.com/articles05/q1/20050324uk-us.shtml).

[21] *VarCon* has also been used for comparing British, American and Canadian English by Cook and Hirst (2012).

[22] For example, a Google search for the first variant pair on the list (*harmonisablest–harmonizablest*) only returns web pages containing different versions of the *VarCon* list.

[23] There are some exceptions, such as *prove*, where the irregular past participle *proven* is the preferred variant in American English (Hundt 1998: 28).

[24] See also Michel et al. (2011: 178), which provides quantitative data on the incidence of regular versus irregular preterites in British and American books 1800–2000, based on data available in Google Books.

[25] The Pearson residual is obtained by subtracting the cell's expected frequency from its observed frequency and dividing the remainder by the square root of the expected

frequency (Agresti 2002: 81).

[26] The shading employs discrete cut-off points at 2 and 4 (–2 and –4), which correspond to residuals individually significant at the α=0.05 and α=0.0001 levels (Meyer et al. 2006: 27).

[27] Due to the overall low frequency of forms, preterites and past participles are not listed separately.

[28] The Chi-squared test is inappropriate if over 20% of cells have expected frequencies of four or less (see e.g. Westfall and Henning 2013: 477).

[29] Accessed on 22 January 2013.

[30] It should also be emphasised that the *WLWC* is based on a snapshot of Wikipedia in April 2010. Accordingly, the results are based on the situation at that specific moment in time and do not reflect any later developments in the contents of the articles.

[31] These samples were obtained using the approach described in Section 5, using lists of cities in these countries as well as the following search words: *Finland*, *Finnish*, *Norway*, *Norwegian*, *Sweden*, and *Swedish*. These three samples are much smaller than Samples 1–3, containing 100, 146, and 143 articles, respectively. The reason is that the English Wikipedia contain fewer articles on these three countries than on the UK and the US, and that articles on cities in the Nordic countries tend to be shorter than 1,000 words (cf. Section 4.1).

[32] As the samples are relatively small, the pair *labor*/*labour* was removed from the list of items. Otherwise, the fact that the Norwegian political party *Arbeiderpartiet* is translated into English as *Labour Party* could potentially skew the results.

Sources

Articles on Wikipedia

7 World Trade Center: http://en.wikipedia.org/wiki/7_World_Trade_Center

Ahmed Raza Khan Barelvi: http://en.wikipedia.org/wiki/Ahmed_Raza_Khan_Barelvi. Early life: http://en.wikipedia.org/wiki/Ahmed_Raza_Khan_Barelvi#Early_life

Conspiracy theory: http://en.wikipedia.org/wiki/Conspiracy_theory

Leeds United A.F.C: http://en.wikipedia.org/wiki/Leeds_United

List of Cities in the United Kingdom: http://en.wikipedia.org/wiki/List_of_cities_in_the_United_Kingdom

List of Towns in England: http://en.wikipedia.org/wiki/List_of_towns_in_England

List of United States cities by population: http://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

Liverpool F.C: http://en.wikipedia.org/wiki/Liverpool_F.C.

Manchester: http://en.wikipedia.org/wiki/Manchester

Neutral point of view: http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

No original research: http://en.wikipedia.org/wiki/Wikipedia:No_original_research

Verifiability: http://en.wikipedia.org/wiki/Wikipedia:Verifiability

Wikipedia: http://en.wikipedia.org/wiki/Wikipedia

Wikipedia's List of Guidelines: http://en.wikipedia.org/wiki/Wikipedia:List_of_guidelines

Wikipedia's List of Policies: http://en.wikipedia.org/wiki/Wikipedia:List_of_policies

Wikipedia's Manual of Style: http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

Other sources

AntConc: http://www.laurenceanthony.net/software.html

Editor Survey 2011: http://commons.wikimedia.org/wiki/File:Editor_Survey_Report_-_April_2011.pdf

Editor Survey 2011/Location & Language: http://meta.wikimedia.org/wiki/Editor_Survey_2011/Location_%26_Language

References

Agresti, Alan. 2002. *Categorical Data Analysis*. Second Edition. Wiley Series in Probability and Statistics. New York: Wiley.

Algeo, John. 2006. *British or American English?: A handbook of word and grammar*

patterns. Cambridge: Cambridge University Press.

Atkinson, Kevin. 2011. "Variant Conversion Info (VarCon) Revision 5.1". http://wordlist.aspell.net/varcon-readme/.

Atwell, Eric et al. 2007. "Which English dominates the World Wide Web, British or American?" *Proceedings of CL2007 Corpus Linguistics Conference*, ed. by Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson. University of Birmingham.

Atwell, Eric et al. 2009. "Arabic and Arab English in the Arab World". *Proceedings of CL2009 International Conference on Corpus Linguistics*, ed by Michaela Mahlberg, Victorina González-Díaz & Catherine Smith. University of Liverpool.

Barbieri, Federica. 2005. "Quotative Use in American English: A Corpus-Based, Cross-Register Comparison". *Journal of English Linguistics* 33(3): 222–256.

Baroni, Marco et al. 2009. "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43(3): 209–226.

Barton, Matt. 2005. *Rhetoric and Composition: A Guide for the College Writer*. Wikibooks.

Barton, Matt & Robert E. Cummings. 2008. *Wiki writing: collaborative learning in the college classroom*. Ann Arbor: University of Michigan Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas et al. 1999. *Longman grammar of spoken and written English*. London: Longman.

Callahan, Ewa S. & Susan C. Herring. 2011. "Cultural bias in Wikipedia content on famous persons". *Journal of the American Society for Information Science and Technology* 62(10): 1899–1915.

Cook, Paul & Grame Hirst. 2012. "Do Web Corpora from Top-Level Domains Represent National Varieties of English?" In: *Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data / 11es Journées Internationales d'Analyse Statistique des Données Textuelles*.

Dalby, Andrew. 2007. "Wikipedias on the language map of the world". *English Today* 23(2): 3–8.

Davies, Mark. 2010. “More than a peephole: Using large and diverse online corpora”. *International Journal of Corpus Linguistics* 15(3): 412.

Denoyer, Ludovic & Patrick Gallinari. 2006. “The Wikipedia XML Corpus”. SIGIR Forum.

Eijkman, Henk. 2010. “Academics and Wikipedia: reframing Web 2.0+ as a disruptor of traditional academic power-knowledge arrangements”. *Campus-Wide Information Systems* 27(3): 173–185.

Elia, Antonella. 2007. “‘Cogitamus ergo sumus’. Web 2.0 Encyclopaedi@s: the case of Wikipedia. A Corpus Based Study”. PhD dissertation, Università degli Studi di Napoli Federico II.

Elia, Antonella. 2009. “Quantitative data and graphics on lexical specificity and index readability: the case of Wikipedia”. *Revista Electrónica de Lingüística Aplicada* 8: 248–271.

Emigh, William & Susan C. Herring. 2005. “Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias”. *Hawaii International Conference on System Sciences*. Los Alamitos: IEEE Press.

Gries, Stefan Th. 2006. “Some Proposals towards a More Rigorous Corpus Linguistics”. *Zeitscrift für Anglistik und Amerikanistik* 54(2): 191–202.

Gries, Stefan Th. 2008. “Dispersions and adjusted frequencies in corpora”. *International Journal of Corpus Linguistics* 13(4): 403–437.

Hinneburg, Alexander et al. 2007. “How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change”. *Literary and Linguistic Computing* 22(2): 137–150.

Hoffmann, Robert. 2008. “A wiki for the life sciences where authorship matters. *Nature Genetics* 40: 1047–1051. doi:10.1038/ng.f.217

Hoffmann, Sebastian. 2007. “Processing Internet-erived Text – Creating a Corpus of Usenet Messages”. *Literary and Linguistic Computing* 22(2): 151–165.

Hundt, Marianne. 1998. *New Zealand English Grammar, Fact or Fiction: A Corpus-based Study in Morphosyntactic Variation*. Amsterdam: John Benjamins.

Hundt, Marianne. 2009. “Colonial lag, colonial innovation, or simply language change?” *One Language, Two Grammars. Differences between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter. Cambridge: Cambridge

University Press, 13–37.

Hundt, Marianne et al. 2007a. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Hundt, Marianne. et al. 2007b. “Corpus linguistics and the web”. *Corpus linguistics and the web*, ed. by Marianne Hundt et al. Amsterdam: Rodopi, 1–6.

Ide, Nancy et al. 2002. “The American National Corpus”. *Proceedings of the 3rd Language Resources and Evaluation Conference LREC*, Canary Islands. Paris: ELRA.

Jucker, Andreas H. 1992. *Social Stylistics: Syntactic Variation in British Newspapers*. Berlin: Walter de Gruyter.

Kehoe, Andrew & Matt Gee 2007. “New corpora from the web: Making web text more ‘text-like’”. *Towards Multimedia in Corpus Studies*, ed. by Päivi Pahta et al. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/02/kehoe_gee/

Kilgarriff, Adam & Gregory Grefenstette. 2003. “Introduction to the special issue on the web as corpus”. *Computational linguistics* 29(3): 333–347.

Kolbitsch, Josef & Hermann Maurer. 2006. “The transformation of the web: How emerging communities shape the information we consume”. *Journal of Universal Computer Science* 12(2): 187–213.

Konieczny, Piotr. 2007. “Wikis and Wikipedia as a teaching tool”. *International Journal of Instructional Technology & Distance Learning*, 4(1).

Konieczny, Piotr. 2014. “Rethinking Wikipedia for the classroom”. *Contexts*, 13(1), 80–83.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington: Center for applied linguistics.

Labov, William et al. 1968. “Empirical foundations for a theory of language change”. *Directions for Historical Linguistics: A Symposium*, ed. by Winfred P. Lehmann & Yakov Malkiel, 95–188. Austin: University of Texas Press.

Larsson, Tove. 2012. “On spelling behavio(u)r: A corpus-based study of advanced EFL learners’ preferred variety of English”. *Nordic Journal of English Studies* 11(3): 127–154.

Levin, Magnus. 2009. “The formation of the preterite and the past participle”. *One Language, Two Grammars. Differences between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 60–85. Cambridge: Cambridge University Press.

Lih, Andrew. 2009. *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. London: Aurum Press.

MacLeod, Donald 2007. "Students marked on writing in Wikipedia". *Guardian Education*, 7 March 2007. https://www.theguardian.com/technology/2007/mar/07/highereducation.elearning

Mair, Christian. 2009a. "British English/American English grammar: Convergence in writing – divergence in speech?" *Anglia - Zeitschrift für englische Philologie* 125: 84–100. doi:10.1515/ANGL.2007.84

Mair, Christian. 2009b. "Corpus linguistics meets sociolinguistics: The role of corpus evidence in the study of sociolinguistic variation and change". *Corpus Linguistics: Refinements and Reassessments*, ed. by Antoinette Renouf & Andrew Kehoe, 7–32. Amsterdam: Rodopi.

Medelyan, Olena et al. 2009. "Mining meaning from Wikipedia". *International Journal of Human-Computer Studies* 67(9): 716–754.

Meyer, David et al. 2006. "The strucplot framework: Visualizing multi-way contingency tables with vcd". *Journal of Statistical Software* 17(3).

Michel, Jean-Baptiste et al. 2011. "Quantitative analysis of culture using millions of digitized books". *Science* 331(6014): 176–182.

Miller, Julia. 2012. "Building academic literacy and research skills by contributing to Wikipedia: A case study at an Australian universitys". *Journal of Academic Language and Learning* 8(2): A72–A86.

Myers, Greg. 2010. *Discourse of Blogs and Wikis*. London: Continuum.

Modiano, Marko. 1993. "American English and higher education in Sweden". *American Studies in Scandinavia* 25(1): 37–47.

Pande, Mani 2011. *Wikipedia Editors Study: Results from the Editor Survey 2011*. Wikimedia Foundation.

Pentzold, Christian & Sebastian Seidenglanz. 2006. "Foucault@ Wiki: first steps towards a conceptual framework for the analysis of Wiki discourses". *WikiSym'06 – Proceedings of the 2006 International Symposium on Wikis*, 59–68. ACM.

Reese, Samuel et al. May 2010. "Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus". *Proceedings of 7th Language Resources and Evaluation Conference LREC'10*, La Valleta, Malta.

Rohdenburg, Günter & Julia Schlüter. 2009a. “Introduction”. *One Language, Two Grammars? Differences between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 1–12. Cambridge: Cambridge University Press.

Rohdenburg, Günter & Julia Schlüter. 2009b. “New departures”. *One Language, Two Grammars? Grammatical Differences between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 364–423. Cambridge: Cambridge University Press.

Romaine, Suzanne. 2008. “Corpus linguistics and sociolinguistics”. *Corpus Linguistics: An International Handbook*, Volume 1, ed. by Anke Lüdeling and Merja Kytö, 96–111. Berlin: Mouton de Gruyter.

Shaoul, Cyrus & Chris Westbury. 2010. *The Westbury Lab Wikipedia Corpus 2010*. Edmonton, AB: University of Alberta. http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tapscott, Don & Anthony D. Williams. 2010. *MacroWikinomics: Rebooting Business and the World*. New York: Portfolio Penguin.

Tardy, Christine M. 2010. “Writing for the world: Wikipedia as an introduction to academic writing”. *English Teaching Forum* 48(1): 12–19.

Walters, Neil L. 2007. “Why you can’t cite Wikipedia in my class”. *Communications of the ACM* 50(9): 15–17.

Walton, Aengus. 2009. *A Statistical Analysis of Stylistics and Homogeneity in the English Wikipedia*. MA Thesis, Trinity College Dublin.

Westfall, Peter H. & Kevin S. S. Henning. *Understanding Advanced Statistical Methods*. Boca Raton: CRC Press.

Widdowson, Henry G. 2000. “On the limitations of linguistics applied”. *Applied Linguistics* 21(1): 3–25.

Young, Jeffrey R. 2006. “Wikipedia founder discourages academic use of his creation”. *The Chronicle of Higher Education*, June 12, 2006. http://www.chronicle.com/blogs/wiredcampus/wikipedia-founder-discourages-academic-use-of-his-creation/2305

Studies in Variation, Contacts and Change in English 15: Texts and Discourses of New Media

Last updated 2023-10-20 by Joe McVeigh