Daniel Schmidtke* and Victor Kuperman

**Mass counts in World Englishes: A corpus linguistic study of noun countability in non-native varieties of English**

DOI 10.1515/cllt-2015-0047

**Abstract:** Research on the morpho-syntax of non-native varieties of English has reported a widespread presence of mass noun pluralization such as *baggages*, *equipments* and *softwares*. In this paper we conducted a corpus linguistic study in order to provide empirically substantiated answers to this claim. We examined the purported prevalence of noun countability in World Englishes in a 1.9 billion-token mega-corpus of global varieties of English. In a comparison of native and non-native varieties of English, first we algorithmically isolated nouns that are more frequently pluralized in non-native varieties. The results indicate a continuum of non-native English countability, along which mass nouns occupy the most extreme tail of inflated occurrences of noun pluralization. In an exploratory analysis, we then examined the similarity in noun countability behaviour across non-native varieties of English. This analysis revealed that geographically proximate countries in which non-native varieties of English are spoken are most similar in the extent to which they pluralize nouns. We argue that noun-countability is a phenomenon best viewed as a gradient that is also regionally dependent.

**Keywords:** World Englishes, morphology, corpus linguistics, count–mass grammar, language variation

1 Introduction

Previous research has proclaimed a notable prevalence of the pluralization of mass nouns (e.g., *informations*, *underwears* and *violences*) in written and spoken language usage of non-native varieties of English (Alsagoff and Ho 1998;
Bautista and Gonzalez 2008; Mollin 2007; Schmied 2008a, 2008b; Schneider 2003b). This is a feature of language production that is manifest in the morphology of non-native speakers of English, where the inflectional plural suffix -s is applied to nouns that are deemed as grammatically ‘mass’ in native English varieties. Scholars cite instantiations of the so-called ‘countable use of mass nouns’ in both written and spoken varieties of English in East Asia, South-East Asia, East Africa and West Africa (Björkman 2008; Crystal 2008; Mesthrie and Bhatt 2008; Platt and Weber 1980). As a consequence of the attention that this morpho-syntactic feature has received in the World Englishes literature (see Hall et al. 2013, for an extensive review), noun countability is declared a discernible linguistic marker of the boundary between native and non-native varieties of English (McArthur 2002; Mesthrie and Bhatt 2008). However, given the paucity of the evidence that goes beyond the anecdotal, it still remains to be seen whether mass noun countability is indeed widespread in non-native Englishes.

The aim of this paper is to provide empirically substantiated results which contribute further insight into the purported prevalence of noun countability in World Englishes and the geographical patterns underlying this variation. The paper is structured as follows. Before arriving at the statistical analyses, we first provide a brief theoretical overview of noun countability in English. Following this, we provide review of the research literature concerning noun countability in World Englishes. Next, with frequency statistics obtained from The Corpus of Global Web-Based English (GloWbE; Davies 2013), a 1.9 billion word corpus of Internet English representing 20 English-speaking countries, we introduce an algorithmic method of quantifying divergent rates of plural form occurrences between native (e.g., Australian, British and Canadian) and non-native (e.g., Ghanaian, Indian and Singaporean) varieties of English. In Study 1 (Section 2) we analyse whether non-native speakers of English do indeed exhibit a tendency to pluralize mass nouns more than native speakers. In Study 2 (Section 3), we then proceed to investigate the patterns of similarity in noun countability preferences across individual non-native varieties of English.

1.1 Noun countability: Theoretical background

Despite the fact that this paper concerns the nature of noun countability in non-native varieties of English, much of the primary linguistic research on noun countability in English has been characterized from the perspective of native-speaker norms. Noun countability refers to the grammatical feature of some languages whereby nouns that form the head of noun phrases are typically either countable or uncountable (also called ‘count’ and ‘mass’). In English,
countable nouns are labelled as such because they can be modified by denumerators and have a morphologically marked plural form (e.g., *one sandwich, two sandwiches). Uncountable nouns are not modified by denumerators, in native varieties of English, but may be modified by ‘vague’ quantifiers such as some (e.g., some advice), and as the default form of the noun is singular, are not morphologically marked for number (e.g., *one advice and *two advices). Thus, the syntactic environments in which one can observe mass nouns are not always straightforward. In a seminal paper, Allan (1980) charted the complex and heterogeneous distribution of quantifiers and nouns in native-speaker English. This research demonstrated that (in native-speaker Englishes) the mass noun admiration can occur with a unit denumerator (e.g., an admiration) but not with a so-called ‘fuzzy denumerator’ (e.g., *several admirations\(^1\)), whereas equipment does not appear with either denumerator types (e.g., *an equipment and *several equipments). Based on this syntactic evidence, Allan (1980) proposed a gradation of ‘noun countability preferences’, which hypothesizes that noun countability is contingent on the syntactic environments in which each noun occurs.

From an alternative perspective, researchers have also attempted to identify a relationship between the non-linguistic semantic properties of noun referents and their linguistically encoded count–mass realizations (Jackendoff 1991; Langacker 2008; Link 1983). The foundational assumption of this approach is that, independently of the linguistic convention used to express number, human beings are able to mentally calibrate the degree to which nominalizable concepts are inherently atomic or not (Bale and Barner 2011; Barner et al. 2009; Imai and Gentner 1997; Wierzbicka 1988). It is proposed that conceptual individuation is granted by the bare inherent semantic properties pertaining to the referent’s qualitative structure, and this is correlated with the noun’s morphosyntactic encoding. For example, count noun concepts are typically bounded and solid, and are denoted by words such as town, pencil and shark. At the opposite end of the same continuum are unbounded objects which are typically amorphous/nonsolid substances or cumulatively mass collections of atomic objects or abstract events, including gravy, grass or advice. However, as Fieder et al. (2014: 2) point out, there is no transparent mapping of the non-linguistic semantic characteristics of mass and count nouns and the way they are morphologically realized. For example, at a conceptual level broccoli and bread are both conceptual entities that are bounded and atomic, yet both are morphologically realized as mass nouns.

\(^1\) We recognize that the asterisks signifying grammatical acceptability are native-speaker centric, yet we preserve them here in order to demonstrate the logic of Allan’s analysis.
In summary, it appears that both syntactic and semantic accounts are each not able to independently predict the count–mass differentiation, and that the count–mass distinction is not strictly binary (Bond and Vatikiotis-Bateson 2002; Kulkarni et al. 2013). Instead, it seems that the reality of the count–mass distinction in natural language usage is probabilistic. That is, all nouns can be inflected with the plural morpheme and this is dependent on semantic and syntactic factors. Thus, one can conceive of a spectrum of countability preference along which nouns are distributed, and is contingent on their semantic and syntactic attributes.

1.2 Noun countability in World Englishes

Another known source of variability in the way in which nouns are pluralized can be traced to the variety of English that exists within the mental lexicon of the individual language user. Research in the World Englishes literature suggests that in non-native English varieties of English, uncountable (i.e., mass) nouns are frequently attested with the inflectional plural suffix -s. A large proportion of the research on mass noun pluralization is qualitative, which has been fruitful in documenting a broad range of English varieties that exhibit the behaviour. For example, Schmied (2008a: 198) reports instances of countable use of mass nouns in Kenyan, Ugandan and Tanzanian Englishes, citing the following usage attested in the East African division of the International Corpus of English (mass noun pluralization underlined):

1. **These advices** are coming because they’ve already studied *all of us*.

Further qualitative research reports additional instances of the countable use of mass nouns in the South-East Asian varieties of English in Malaysia, the Philippines and Singapore (Alsagoff and Ho 1998; Bautista and Gonzalez 2008; Platt et al. 1984; Schneider 2003b). For example, Gonzalez (1983: 167) cited in Cane (1994) reports the following usage attested in the East African division of the International Corpus of English (countable usage underlined):

2. **He has many luggages**.

Moreover, Cane (1994: 354) reports the following example as “instability of singular/plural countable usage” in Brunei English:

3. **And here’s an advice** for you all.

Quantitative attempts to measure the extent and variability of noun countability across English varieties are more scarcely found in the research literature.
For example, Mollin (2007) conducted a corpus linguistic analysis of non-Anglophone European English. In this research, Mollin queried a 400,000 word corpus of transcribed public speeches and written online chatroom discussions from the European Commission, representing a total of 900 different speakers. The low validity of the results afforded by the small corpus size and narrow scope of representative language genres notwithstanding, Mollin reported a significant overrepresentation of pluralized nouns such as evidences and advices, yet still found that overall usage of mass nouns was negligible. From this, Mollin concluded that in contexts where English is used as foreign language, pluralization of mass nouns is a significant but not prominent feature.

Mollin’s findings dovetail with further corpus linguistic findings reported by Hall et al. (2013), who harnessed the World Wide Web as a language corpus to quantify the countable usage of mass nouns across a variety of Englishes from around the globe. This study was the first major corpus linguistic attempt to gauge the extent of noun countability across varieties of English. Using Google’s ‘advanced’ search engine tool to collect frequency counts of countable uses of mass nouns, Hall et al. (2013) compared mass noun pluralization of British English users to that of non-native English users from 14 different non-native English-speaking countries (e.g., Bangladesh, Hong Kong, Iceland & Mexico). Using the novel research methodology of exploiting the domain-specific Google search bar to target a particular variety of English used by a country (e.g., .ph identified the webpages of the Philippines, .hk identified the webpages of Hong Kong, .uk targeted the web pages of the United Kingdom, and so on), they queried Google with a list of 25 ‘truly uncountable’ mass nouns (see list below). These nouns were selected based on a syntactically informed ‘noun countability preference’ criteria (Allan 1980) that were filtered out from a larger set of 215 mass nouns. Hall et al.’s (2013) list of 25 search terms were as follows; advice, applause, baggage, cash, corruption, dew, employment, equipment, evidence, feedback, fun, furniture, hardware, homework, information, jewellery, knowledge, luck, luggage, magic, slang, software, traffic, underwear and violence.

In this research, Hall et al. adopted the influential Three Circles model of World Englishes (Kachru 1990, 1992) in order to classify the sample of Englishes into three broad groups (or circles). We also adopt the terminology and system of classification of The Three Circles model in the current study. The Inner Circle (IC) of English is represented by the speakers of countries that use English as a Native Language (e.g., Australia, Britain, Canada, New Zealand and the United States). The Outer Circle (OC) is comprised of Second Language (L2) English users who inhabit post-colonial countries, such as Bangladesh, Kenya, India, the Philippines and Singapore. For these countries, English is institutionalized and serves as a language of governmental policy, administration, jurisdiction and media
broadcasting (Schneider 2003a). Finally, the Expanding Circle (EC), which is a circle of Englishes not addressed in the current study, represents the English spoken by populations of countries such as Angola, China, Iceland, Mexico and China. This is where English does not function as an official language, but instead prevails as a ‘Foreign Language’ in tertiary education, international governance, the media and also in the arenas of science and commerce. It should be noted that the aforementioned description is oversimplified: EC varieties of English also include speakers who are first and second language users.

Hall et al. (2013) reported a significant difference in the average proportion of pluralized mass noun frequencies between British English (0.01%) and EC and OC English usage combined (2.22%). This finding was consistent when comparing combined OC and EC frequencies with both Google searches and frequency counts from the British National Corpus. Moreover, the difference between the average mass noun pluralization frequencies of speakers of OC (3.43%) and speakers of EC English (1.01%) was also statistically significant. In addition, they reported that these aggregated rates were not influenced by the higher pluralization rate of a single noun, but that a range of 18 different nouns out of the list of 25 contributed to the highest levels of noun countability and were differentially distributed across each country.

To sum up, a majority of the evidence that points towards the pluralization of mass nouns in non-native English varieties is qualitative. The merit of this research lies in the very discovery of this linguistic feature, and the supporting examples of countably used mass nouns from a range of English varieties. However, aside from the two corpus studies that we have mentioned, little is known about the precise extent of the countable usage of mass nouns. In what follows we set out to empirically scrutinize the findings of the research summarized above, which suggests mass noun countability in Outer Circle varieties of English occurs to a greater extent in comparison with the Inner Circle. In our first analysis we thus explored the extent of noun countability in OC varieties of English.

2 Study 1: Detecting mass noun pluralization in non-native varieties of English

This analysis aims to estimate the scope of mass noun pluralization in all of the Outer Circle varieties of English combined. The motivation behind this investigation was to examine the purported asymmetry in the magnitude of pluralization of mass nouns across native and non-native varieties of English, and test
whether their is indeed a good reason to be surprised by the frequency of occurrence of mass noun pluralization in Outer Circle Englishes.

In order to maximize the retrieval of pluralized mass nouns in the OC, our method adopted a combination of bottom-up and top-down analytical procedures. We first employed a bottom-up procedure to retrieve the greatest possible number of nouns, irrespective of their count–mass identity, that are significantly more pluralized in the OC relative to the IC. Once we identified the pool of nouns and their singular and plural frequencies, we applied a statistical metric to our data set, which analyses the divergence in the frequencies of plural forms between IC and OC Englishes. Using frequency of occurrences of singular and plural forms of nouns as a dependent measure, the test statistic provides a robust and precise method of identifying the pluralized nouns that maximally discriminate native and non-native varieties of English. Next, we tested the dispersion of known ‘mass’ nouns within the continuum of countability that is generated by the aforementioned frequency discrimination test. The logic is that if ‘mass’ nouns tend to be clustered among the most extreme end of the continuum, where pluralization is greatest in the OC overall, then this would imply that mass nouns are more pluralized than count nouns in the OC.

The reasons behind remaining blind to the count–mass distinction in the initial step of the exploration were both methodologically and theoretically motivated. Because the corpus we used is part-of-speech (POS) tagged, we were able to employ a search technique that was not restrictive. That is, we were able to objectively collect plural frequencies of all possible unique nouns in the corpus. The advantage of this approach is that it forced us to remain initially neutral to the count–mass distinction and also examine mass noun pluralization within the context of as many nouns available in the language. In addition, it alleviated any theoretical limitations that may have been imposed by performing a rudimentary search in the corpus for a small, pre-specified and potentially incomplete list of mass nouns.

It is necessary to stress at this point that we are aware that large-scale quantitative research often abstracts away from the qualitative detail, such as the contextual and syntactic underpinnings of the phenomenon under critical examination. Thus, the research presented here only complements and does not supercede other work on noun countability in World Englishes. However, we would also argue that the current digital age allows us to perform a macro-level sweep of linguistic phenomena, and contemporary computational linguistic tools allow us to do so with a higher level of quantitative sophistication. Thus, a key advantage of the relatively recent introduction of mega-corpora and a slew of innovative corpus linguistic techniques is that the researcher can now capture large-scale global trends of language usage.
2.1 Methods

2.1.1 The GloWbE corpus

We refined elements of the corpus-oriented research design and methodology used by Hall et al. (2013). The Google search tool permitted Hall et al. to harvest a broad snapshot of variability in the usage of lexical forms from many country domains. Despite this innovation, the cost of this approach is that language data extracted from raw Internet searches can be noisy and intractable. For example, the Google search technique returns raw unfiltered data, with no guarantee that the author(s) of the web page are citizens of the country-specific domain under which the text is searched. Also under question is the issue of text authenticity. This is problematic because a web page’s text could be duplicated or produced by an auto-generated spam page, thus inflating word frequency counts. Moreover, perhaps the most restrictive element of their methodology was that Hall et al.’s search query was limited to just 25 nouns and many potential mass nouns were excluded because they share the same orthographic form as other parts of speech, such as research/researches which could be two parts of speech, either a noun or verb.

Many of the aforementioned methodological drawbacks of Hall et al.’s research technique were remedied with the recent publication of GloWbE (Davies 2013; Davies and Fuchs 2015). The corpus incorporates Google’s web page data to better pinpoint the provenance of the author of the web page. For example, if the domain address is.com, then the geographical origin of the web page can be determined from the IP address of the web page, and if still uncertain, approximates authorship location using the IP information of the page’s incoming traffic and of the links to the webpage. It is true that these measures do not certify the linguistic identity of an author’s text completely (Mukherjee 2015; Nelson 2015), yet this is certainly an advancement upon Hall et al.’s method. Moreover, Davies (2013) notes that the data-mining software used to construct GloWbE minimizes inter and intra web page text duplication. Finally, of crucial importance to the current study, GloWbE is POS tagged, which meant that we could search specifically for instances of nouns.

Despite the faithfulness of GloWbE to the true plurilithic reality of World Englishes, there are still drawbacks that one must consider when using GloWbE as tool to examine the structure of World Englishes. The first is the amount of “informal and nonstandard” usage that is introduced by the relatively unsanitized web-crawling method (Mair 2015: 30). Although these “nonstandard” instances of linguistic production are precisely the object of importance to many researchers, they pose problems for POS tagging tools (for further discussion see Mair 2015).
Moreover, the unrestricted sampling of blog and website material also inflates the probability of occurrences of “unintelligible and uninterpretable gibberish” (Mair 2015: 31). As a consequence of such noise and intractability, as is also suggested by Nelson (2015: 39), we advocate using GloWbE in conjunction with more “carefully-curated” language corpora (see Data collection below).

We sampled 6 IC Englishes and 12 OC Englishes from the GloWbE corpus (see Table 1 for summary of Inner and Outer Circle breakdown). Of the total of 20 countries represented in GloWbE, we omitted South Africa and Jamaica as they proved to be difficult to categorize. For example, the national boundaries of South Africa enclose at least three notable varieties of English (White South African English, WhSAfE; Indian South African English, InSAfE and Black South African English, BISaFe), of which WhSAfE is classified as a native variety whereas InSAfE and BISaFe are classified as non-native varieties. We also omitted Jamaica; while English is the official language of Jamaica, there are wide variety of English, from Caribbean Standard English to pidgins and creoles. Researchers (see Jenkins 2003) raise concerns that these characteristics render it especially difficult to classify Jamaica as an Inner Circle English. Lastly, while the reason for not including EC varieties in our analysis is due to their absence

<table>
<thead>
<tr>
<th>Kachruvian Circle</th>
<th>Country</th>
<th>Total number of words in GloWbE</th>
<th>n noun types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner Circle</td>
<td>Australia</td>
<td>148,208,169</td>
<td>8,048</td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>134,765,381</td>
<td>8,017</td>
</tr>
<tr>
<td></td>
<td>Great Britain</td>
<td>387,615,074</td>
<td>9,188</td>
</tr>
<tr>
<td></td>
<td>Ireland</td>
<td>101,029,231</td>
<td>7,599</td>
</tr>
<tr>
<td></td>
<td>New Zealand</td>
<td>81,390,476</td>
<td>7,194</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>386,809,355</td>
<td>9,105</td>
</tr>
<tr>
<td>Outer Circle</td>
<td>Bangladesh</td>
<td>39,658,255</td>
<td>5,938</td>
</tr>
<tr>
<td></td>
<td>Ghana</td>
<td>38,768,231</td>
<td>5,474</td>
</tr>
<tr>
<td></td>
<td>Hong Kong</td>
<td>40,450,291</td>
<td>6,424</td>
</tr>
<tr>
<td></td>
<td>India</td>
<td>96,430,888</td>
<td>7,107</td>
</tr>
<tr>
<td></td>
<td>Kenya</td>
<td>41,069,085</td>
<td>5,537</td>
</tr>
<tr>
<td></td>
<td>Malaysia</td>
<td>42,420,168</td>
<td>6,090</td>
</tr>
<tr>
<td></td>
<td>Nigeria</td>
<td>42,646,098</td>
<td>5,761</td>
</tr>
<tr>
<td></td>
<td>Pakistan</td>
<td>51,367,152</td>
<td>5,847</td>
</tr>
<tr>
<td></td>
<td>Philippines</td>
<td>43,250,093</td>
<td>6,554</td>
</tr>
<tr>
<td></td>
<td>Singapore</td>
<td>42,974,705</td>
<td>6,087</td>
</tr>
<tr>
<td></td>
<td>Sri Lanka</td>
<td>46,583,115</td>
<td>6,033</td>
</tr>
<tr>
<td></td>
<td>Tanzania</td>
<td>35,169,042</td>
<td>5,422</td>
</tr>
</tbody>
</table>
in GloWbE, it is also worth adding that the distinction between OC and EC varieties is far from straightforward and is a contentious issue within the World Englishes literature (e.g., see Seidlhofer 2004). Thus, by not including EC varieties in our sample of global varieties of English, we control for the additional layer of complexity that this fuzzy distinction would have introduced to the study.

2.1.2 Data collection

We began by obtaining an exhaustive list of nouns from SUBTLEX-US, a 51-million-word corpus of US film and media subtitles (Brysbaert and New 2009). We used SUBTLEX-US (Brysbaert et al. 2012), instead of directly drawing our items from GloWbE, because SUBTLEX-US’s POS information allowed us to obtain noun lemma forms that were automatically identified as a noun in 94% or more of their occurrences. This ensured that we collected frequency estimates for nouns that are highly likely to be a noun, e.g., *abdomen*, and that we do not consider nouns such as *square*, which is used a noun in only 50% of its occurrences according to SUBTLEX-US. This step created a list of 17,757 noun lemmas, which were then queried in GloWbE, ensuring separate frequencies were obtained for singular and plural forms of each noun. This search returned 9,746 nouns that occurred in the singular and plural form at least once in any country-specific domain in GloWbE. We then aggregated the frequency counts of all singular and plural forms separately, for both the IC and the OC. The resulting dataset consists of a list of nouns each with a corresponding IC singular form frequency, OC singular form frequency, IC plural form frequency and OC plural form frequency. For example, the word *abattoir* occurs 504 times in the singular form in the IC and 123 times in the singular form in the OC. The plural form, *abattoirs*, occurs 276 times in the IC, and 94 times in the OC.

To control for potential typographic errors and misspellings, we also removed words that occupied the bottom 25%, separately, of (1) the joint singular and plural frequency distribution of the IC and (2) the joint singular and plural frequency distribution of the OC. This operation equated to removing all nouns with a joint singular and plural frequency of below 367 and 121 occurrences from the respective IC and OC frequency distributions. Our data clean-up procedure also included the removal of 84 nouns that were attested with two plural forms, such as *angel* which has two separate plural spelling formats, *angels* and *angeli*. We also removed 608 nouns with British and US spelling variants (for both singular and plural forms), such as *colour/color*. British spelling formats were
obtained from the SUBTLEX-UK corpus (van Heuven et al. 2014). These clean-up steps reduced our data pool to a total of 6,227 unique nouns, each with separate frequency counts for their singular and plural forms in both the Inner and Outer circle English usage.

2.1.3 Statistical analyses

With the resulting pool of 6,227 nouns, we singled out the nouns that are statistically overrepresented in the plural form in the Outer Circle relative to the Inner Circle. To achieve this, we applied the log-odds ratio informative Dirichlet prior (LORIDP) method as used by Monroe et al. (2008) (see also Jurafsky et al. 2014). This method was proposed in order to find words that are statistically overrepresented in a particular category of documents compared to another. In the present case, the two documents under comparison are the Outer and Inner Circle Englishes of GloWbE. The LORIDP method estimated the difference between the frequency of the plural form of noun $w$ in the respective Outer Circle ($i$), and Inner Circle ($j$), subdivisions of the GloWbE corpus via the log-odds ratio for $w$, $\delta^{(i-j)}_w$, which is computed as

$$\delta^{(i-j)}_w = \log \left( \frac{y^i_w + \alpha_w}{n_i + \alpha_0 - (y^i_w + \alpha_w)} \right) - \log \left( \frac{y^j_w + \alpha_w}{n_j + \alpha_0 - (y^j_w + \alpha_w)} \right),$$

where in this case, $n_i$ is the total number of noun tokens in the Inner Circle subdivision of the GloWbE corpus $i$, $n_j$ is the total number of noun tokens in the Outer Circle subdivision of the GloWbE corpus $j$, $y^i_w$ is the count of the plural form of noun $w$ in corpus $i$, $y^j_w$ is the count of the plural form of noun $w$ in corpus $j$, $\alpha_0$ is the total number of tokens in the GloWbE corpus and $\alpha_w$ is the total frequency of the singular and plural forms of noun $w$ in all of GloWbE. The variance of the above measure was then estimated as

$$\sigma^2(\delta^{(i-j)}_w) \approx \frac{1}{y^i_w + \alpha_w} + \frac{1}{y^j_w + \alpha_w},$$

Lastly, the z-score statistic of the LORIDP of each noun was calculated as

$$\frac{\delta^{(i-j)}_w}{\sqrt{\sigma^2(\delta^{(i-j)}_w)}}.$$

The main advantage of the LORIDP measure is that it shrinks the prior probability of word frequency by using counts from a representative background corpus (in this case $\alpha_w$). This approach allows differences to be detected even between the highest
frequency plural forms across the OC and IC, a restriction not afforded by the $\chi^2$ test of non-homogeneity of variance and several other measures of association strength. Another advantage of this method is that by using the $z$-score of the log-odds ratio, the amount of variance in word’s frequency across the Inner and Outer Circle is controlled for (see Gries 2014, for an overview of other available measures of statistical association for word frequency data and Appendix A for a comparison of LORIDP with two competing difference co-efficients).

The resulting range of 6,227 LORIDP $z$-scores represents a standardized measurement of the divergence in noun pluralization frequencies between Inner and Outer Circle Englishes. A noun associated with a negative value indicates a degree of ‘overrepresentation’ of its plural form in Outer Circle Englishes, and a noun associated with a positive value indicates a degree of ‘overrepresentation’ of its plural form in Inner Circle Englishes. If the plural form of a noun is equally well represented in both the Inner and Outer Circle, the formula yields a $z$-score of exactly 0.00. For example, the noun information has the most negative $z$-score ($z = -113.69$) and represents the furthest extreme of Outer Circle pluralization. Conversely, thing ($z = 491.22$) lies at the opposite end of the scale of countability, which is instead occupied by nouns that are more likely to be pluralized by Inner Circle English users.

Our research question concerns the nouns that are more frequently pluralized in the Outer Circle. We thus used the LORIDP $z$-score scale to discern the nouns for which the frequency of the plural form was significantly greater in the Outer Circle as compared to the Inner Circle. We set out to isolate the nouns with a one-tailed significance at the probability threshold of $p \leq 0.01$. Note that in the present scenario, the LORIDP measure calculated the divergence in plural usage between the Inner and Outer Circle for a multitude (6227) of nouns. The inflated Type I error rate resulting from the consideration of multiple comparisons was thus handled with the application of the Bonferroni correction. After the application of the Bonferroni correction ($\alpha = 0.01/6227$), this operation involved identifying the nouns with a $z$-score less than $-4.66$ (i.e., isolating nouns less than 4.66 standard deviations below the mean LORIDP score). This analysis produced a pool of 1,591 nouns that are significantly more pluralized by Outer Circle speakers of English relative to Inner Circle speakers of English.

Next, we combined top-down theoretical linguistic knowledge with the LORIDP metric by analysing the dispersion of a pre-defined list of mass nouns within the full order of nouns ranked by the LORIDP metric. A list of 74 mass nouns were taken from Hall et al.’s (2013) materials and were matched with nouns that were present in our total list of 6,227 nouns. These materials were selected such that they were ‘truly uncountable’ according to Allan’s (1980) countability preferences criteria (this list of nouns was used by Hall et al. 2013).
If these pre-specified mass nouns are more likely to be associated with $z$-values lower than the significance threshold of $-4.66$ (i.e., are significantly more likely to be pluralized in the OC than the IC), then this would confirm that mass nouns are most likely to be used countably by OC users of English. Thus, the LORDIP measure would successfully validate the claim that OC English speakers more frequently pluralize mass nouns. As a consequence, this finding would also justify the use of the LORIDP measure in order to consider the remainder of the nouns that also inhabit this end of the scale as additional countably used mass nouns. Alternatively, if there is no observed pattern of the dispersion of mass nouns across this continuum, then this would entail that mass nouns do not tend to be pluralized more in the OC.

2.2 Results and discussion

A non-random pattern was observed, such that the nouns from the pre-selected mass noun list were more prevalent among the set of nouns that are statistically more pluralized in the OC relative to the IC. We found that 73 of the 74 pre-defined mass nouns were among the 1,591 nouns that were significantly more frequently pluralized in the Outer Circle (the only mass noun more frequently pluralized by native speakers relative to non-native speakers was *permissions*). The observed pattern was confirmed statistically in a $\chi^2$ test of independence (goodness-of-fit test) which tested the relationship between the frequencies of mass nouns that were present on either side of the LORIDP significance threshold. As stated earlier, there were 1,591 nouns that were significantly more frequently pluralized in the Outer Circle. Thus, we compared the total number of mass nouns present within this pool of 1,591 words, with the total number of mass nouns found within the 3,659 nouns that are significantly more likely to be pluralized in the Inner Circle, i.e., the number of nouns with a LORIDP $z$-score of greater than 4.66. As a result of the imbalance in the number of nouns at either side of the significance threshold, we ensured that the $\chi^2$ test accounted for the difference in expected frequencies of mass nouns in either pool of words (Table 2 reports observed and expected frequencies). The $\chi^2$ test confirmed that the number of pre-defined mass nouns was statistically overrepresented within the list of nouns that are pluralized more frequently by Outer Circle speakers ($\chi^2 (df = 1, n = 73) = 163.65, p \leq 0.0001$).

Furthermore, Figure 1 also suggests a gradation in the concentration of the pre-selected mass nouns as pluralization in the OC increases. More specifically,

---

2 We opted to use the $\chi^2$ goodness-of-fit test instead of Fisher’s exact test because the expected frequency counts are greater the 5. For completeness we also ran Fisher’s exact test, which revealed that the imbalance in frequencies are significantly different from expected.
Table 2: A table displaying the expected and observed frequencies of mass nouns which were found within the pools of significantly pluralized nouns of the Outer and Inner Circle.

<table>
<thead>
<tr>
<th></th>
<th>Outer Circle</th>
<th>Inner Circle</th>
</tr>
</thead>
<tbody>
<tr>
<td>n nouns attested more frequently in plural form</td>
<td>1,591</td>
<td>3,659</td>
</tr>
<tr>
<td>n mass nouns expected</td>
<td>22</td>
<td>52</td>
</tr>
<tr>
<td>n mass nouns observed</td>
<td>73</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1: A dispersion plot of 73 mass nouns (grey lines) distributed across the range of \( z \)-scores of the log-odds ratio informative Dirichlet prior (LORIDP) measure. The dashed line depicts the significance threshold whereby all 1,591 nouns to left of the line are significantly more pluralized in the Outer Circle relative to the Inner Circle \((p \leq 0.01)\). The dotted line highlights the \( z \)-score equal to 0 (i.e., where the rate of pluralization is equivalent in Inner and Outer Circle Englishes). The y-axis displays a selection of 37 out of the total 73 mass nouns (to preserve legibility) and are ranked in descending order of the LORIDP \( z \)-score.
the plot suggests an uneven dispersion of mass nouns along the LORIDP continuum such that mass nouns gradually become more prevalent as LORIDP z-scores become more negative (more frequently pluralized in the Outer circle). We tested this observation statistically by binning the 1,591 nouns (ranked by the LORIDP z-score) into deciles and then counting the number of predefined mass nouns in each decile (numerically larger deciles contain groups of nouns that are relatively less frequently pluralized in the OC). Spearman’s $\rho$ revealed a significant negative correlation between the number of mass nouns in each decile and the ordinal number of the decile (1 to 10), $r_s = -0.86$, $p < 0.0001$. Figure 2 visualizes the counts of mass nouns per each decile of the LORIDP test.

![Figure 2: A bar plot of the relative frequencies of mass nouns within the 1,591 nouns that were significantly more frequently pluralized by Outer Circle users in GloWbE. Counts of mass nouns were summed per decile of the ranked order of LORIDP z-scores (lower deciles indicate greater OC pluralization). Absolute frequencies are plotted above each bar.](image)

The results revealed a selection of pluralized mass nouns such as awarenesses, educations, infrastructures, migrations and entertainments. These nouns were not
part of the pre-specified list of mass nouns. Examples of these nouns used in context from various sections of GloWbE are given below (countable usage underlined):

4. *Whereas quiet awarenesses, perceptions and intuitions stimulate the heart...* (India, G)

5. *... factors like lack of educations, infrastructures,... are also among the important causes of rural migrations...* (Kenya, G)

6. *... if people are poor their children get bad educations...* (Hong Kong, G)

7. *The large amphitheatre must have provided entertainments to the residents.* (Singapore, G)

8. *The British love entertainments of all kinds.* (Sri Lanka, B)

These nouns complement the list of pluralized mass nouns that have been identified in prior research, such as *advices* (Schmied 2008b), *luggages* (Gonzalez 1983) and *informations* (Hall et al. 2013). Examples of these usages in context are provided below:

9. *My wise shaykh the murshid, Shihab, gave me two advices...* (Pakistan, G)

10. *The first branch closed early, and with three luggages in my hands...* (Philippines, B)

11. *So you are hereby advised once more to keep your informations strictly confidential...* (Ghana, B)

Taken together, our results indicate that mass nouns are most likely to be more frequently pluralized by non-native speakers of English. Moreover, the overall pattern suggests that pre-selected mass nouns gradually become more densely concentrated as the degree of pluralization in the OC increases. The full list of 1,591 nouns (with a corresponding LORIDP z-score and ratio signifying divergence in the rates of pluralization between the IC and OC) is provided as a supplementary data file.

3 Examples in GloWbE are referred to with the country of origin. The third letter refers to text type, with G standing for “General (may also include blogs)” and B for “(only) Blogs”, accessed 9 November 2015.
3 Study 2: Patterns of pluralization within non-native varieties of English

The results of the Study 1 support the findings of previous research which suggests that mass nouns are more frequently pluralized by non-native speakers of English. In Study 2, we conducted an exploratory investigation on the similarity of noun pluralization frequencies between individual OC countries represented in GloWbE. Our hypothesis was guided by Nerbonne’s Fundamental Dialectal Postulate (Nerbonne and Kleiweg 2007: 157), which states that “Geographically proximate varieties tend to be more similar than distant ones” and Tobler’s first law of geography (Tobler 1970: 237), which states that “Everything is related to everything else, but near things are more related than distant things”. Thus, the central assumption that we adopt is that geography to some extent influences dialectal variation (see also Séguy 1971). In this analysis, we hypothesise that geographically closer varieties of English have a higher probability of contact, and correspondingly have a higher probability of pluralizing the same set of nouns with similar proportional usage.

3.1 Methods

We selected the set of 1,591 nouns that were identified as significantly more pluralized in Outer Circle English in Study 1. These are a subset of nouns where the rate of OC pluralization is both maximally different from the IC and, also where the highest proportion of mass nouns are found. With these nouns we aggregated their singular and plural frequency counts per each of the 12 countries representing the OC in GloWbE. We then applied the LORIDP method to iteratively compare the rate of pluralization of each individual OC country with the aggregated plural and singular frequencies of the remaining 11 OC countries. The LORIDP algorithm was the same as defined in Study 1, but in this case, \( n_j \) is the total number of noun tokens in the \( n \)th country in the Outer Circle of the GloWbE corpus \( j \), \( n_i \) is the total number of noun tokens in the aggregated Outer Circle GloWbE corpus \( i \) (excluding the plural and singular counts from country \( j \)), \( y'_w \) is the count of the plural form of noun \( w \) in corpus \( i \), \( y'_w \) is the count of the plural form of noun \( w \) in corpus \( j \), \( \alpha_0 \) is the total number of tokens in the GloWbE corpus and \( \alpha_w \) is the total frequency of the singular and plural forms of noun \( w \) in all of the Outer Circle.

The resulting data source consisted of a data frame (1591 nouns x 12 countries) wherein each cell contained a \( z \)-score of the LORIDP metric. For
example, the cell for the noun *remainder* in Bangladesh contained a z-score of 0.06. This indicates that for this specific noun, there is little difference in the proportional frequency of the plural form of this noun in Bangladesh in comparison with the rest of the OC countries combined. With this data set, we then were able to assess the (dis)similarity of noun pluralization usage among the 12 countries in the OC. To do this, we calculated the Euclidian distances over each possible pair of the 12 vectors (n countries) containing 1,591 (n nouns) z-scores. The outcome of this process is a 12 × 12 dissimilarity matrix of pluralization across the countries of the OC represented in GloWbE.

Next, we used Ward’s Method of agglomerative clustering technique (Ward 1963) to build a hierarchy of clusters based on the set of pluralization distances between each country. In Ward’s clustering process, each element (country) is clustered using Ward’s minimum variance method, which merges pairs of clusters based on the pair of elements that yield the smallest error sum of squares. The clusters with the smallest variance between two elements are fused together to form larger clusters. This process occurs sequentially until each cluster is linked in a hierarchical fashion. Crucially, this method is agnostic to the geographical location or typological proximity of the L1 substrates of each of the 12 OC countries, and simply works to find the optimal set of clusters based only on the relative magnitude of noun pluralization across each country. Uncertainty in the hierarchical clustering analysis was computed using the pvclust (Suzuki and Shimodaira 2006) package in the R statistical computing software program (R Core Team 2014). This process computes a p-value (approximately unbiased) for each cluster, which indicates how strong the cluster is supported by data.

### 3.2 Results and discussion

The visualization of the hierarchical cluster analysis in Figure 3 reveals an astonishing relationship between the geographical location of each country and the extent to which each country pluralizes nouns in English. The dendrogram shows a visually impressive recreation of geographical proximity from (dis)similarities in the frequency with which OC countries pluralize nouns. The highest split in the dendrogram separates three South Asian countries (Pakistan, India and Sri Lanka) from the rest of the OC countries. Within this branch, Pakistan splits from the sub-branch of India and Sri Lanka. The remaining branch of OC countries is split into a series of nested clusters which first bifurcates all four African countries from the remaining countries. Within this cluster, geographical convergence is clear with the presence of the paired cluster of Kenya and Tanzania. Kenya and Tanzania are both East African countries that
share a political border. Furthermore, Nigeria and Ghana, who represent the highest split within African subdivision of the dendrogram, are also within close geographical proximity of one another in Western Africa (but are not contiguous countries). The remainder of the dendrogram visualizes a separation of all four Southeast and East Asian countries from the African countries. Within this cluster, Hong Kong (East Asia) splits from The Philippines, Malaysia and Singapore (Southeast Asia). Malaysia and Singapore are the most closely related in this cluster, which is striking since they are both closest in geographical space and also both share an L1. The obvious incongruency in the dendrogram is the fusion of Bangladesh with the East and Southeast Asian countries. It is not immediately clear exactly why this is the case. Finally, we report that the multiscale bootstrap-based \( p \)-value reported above each split reveals that the cluster classification is not extremely stable, so we urge the reader to treat the clusters with caution. However, it is still clear that the broad pattern of the

Figure 3: A dendrogram of selected OC countries in the GloWbE corpus clustered based on the relative magnitude of pluralization of 1,591 nouns that are pluralized more frequently in the OC as a whole. The vertical distances between cluster splits are proportional to the square of the Euclidian distance of the LORIDP metric. Clustering was estimated using Ward’s method. The multiscale bootstrap-based \( p \)-values for each cluster are displayed below the split in the branch of the dendrogram.
country-level hierarchical clustering analysis, which is based purely on the (dis)similarity of the relative frequency of noun pluralization, converges with the geographical distribution of varieties of English.

4 General discussion

The primary aim of Study 1 was to verify the claims that Outer Circle users of English pluralize nouns that are semantically and grammatically defined as ‘mass’. The aim of Study 2 was to explore the degree of similarity among OC countries in their scope of noun pluralization. This section will discuss the findings of each research objective in turn.

4.1 Detecting the nouns that count in Outer Circle Englishes

The results of Study 1 provide solid empirical evidence that across the range of all nouns, mass nouns (according to IC grammars) are most likely to be pluralized by non-native speakers of English. More precisely, we revealed the novel finding that ‘truly uncountable’ nouns gradually become more numerous at the extreme tail of the LORIDP continuum of noun countability, where nouns are attested more frequently with plural morphology in the OC relative to the IC. These findings both refute the null hypothesis that all count and mass nouns are pluralized proportionally more in the OC compared to the IC.

The metric we used to assess the rate of OC countability was blind to the theoretically grounded grammatical count–mass distinction, but entered into a systematic convergence with a subset of nouns encoded with this grammatical information. The \(\chi^2\) test in Study 1, confirmed the existence of a non-random dispersion of 74 mass nouns along the LORIDP continuum, such that mass nouns (at least the ones from our non-exhaustive list) are more likely to be pluralized by OC speakers. This finding provides robust evidence that, relative to the varieties of Inner Circle English that we sampled, mass nouns tend to be pluralized more frequently by Outer Circle English speakers. Moreover, the Spearman rank correlation coefficient in Study 1 indicated that the concentration of mass nouns becomes more densely compact as the degree of OC pluralization increases. The detection of a gradual increase in the presence of mass nouns as the degree of OC pluralization increases suggests that the phenomenon that we set out to explore may be more subtly conceptualized as a ‘continuum of non-native English mass noun pluralization’.

Daniel Schmidtke and Victor Kuperman

Authenticated | schmiddf@mcmaster.ca author's copy
Download Date | 1/13/16 7:45 AM
In addition, we argue that one can use these statistically supported observations as justification for using the LORIDP test to explore the nature of the mass-count characteristics of nouns that occupy the end of continuum which mass nouns are most likely to inhabit. As a result, we were also able to confirm the usage of a wealth of previously unidentified ‘mass’ nouns which were used countably by non-native speakers, such as *moneys, educations, attentions* and *trainings*. We also reconfirmed the results of studies (Björkman 2008; Crystal 2008; Gonzalez 1983; Hall et al. 2013) that previously reported the countable use of mass nouns such as *advices, baggages, furnitures, informations, slangs* and *softwares*.

As we stressed earlier, the initial sampling method used to pinpoint proportional differences between the extent of pluralization across the IC and OC did not take into account the count–mass distinction. We found that this method offered a number of advantages to our research aim. By remaining unguided by the assumptions of linguistic theory at the initial collection phase, we were able to maximize the identification of a range of grammatically mass nouns out of the full range of nouns available to use in the language corpus. Secondly, once combined with a top-down approach, the log-odds ratio informative Dirichlet prior measure enabled us to establish a criterion with which to systematically isolate, with statistically supported confidence, a list of nouns that are pluralized significantly more frequently by OC English users from within the full range of nouns available in the language.

### 4.2 Intra-OC variability in noun pluralization

In Study 2, we adopted an agglomerative hierarchical clustering technique to trace groups of countries based on the individual country-level frequency estimates of the plural forms of 1,591 nouns uncovered in Study 1. Based on this analysis, we found the non-random pattern that geographically proximate countries share similar lexical usage. Why do we observe this pattern of results? We must stress that our data set did not allow us to go beyond national-level geography in order to pinpoint the many underlying factors that could potentially govern the observed patterns. Indeed, geography itself can be construed as a distal cause of the observed structure of the hierarchical cluster analysis. We believe that in order to reach a more precise understanding of the factors underlying noun pluralization more work must be done to disentangle historical factors (e.g., colonial history) underlying the borrowing and appropriation of emergent linguistic conventions (Schneider 2010), genealogical factors that influence the processing of linguistic nativization (Cysouw 2013) and also,
from a psycholinguistic perspective, the cognitive process of cross-linguistic transfer via the L1 substrate (Hall et al. 2013). Each of these factors was not encoded in our data set, but theoretically they could each contribute individual amounts of variability to the observed patterns that geographically proximate countries share similar lexical usage.

To take a geographical perspective, the patterns in our results provide concrete support for Leitner’s (1992) account of English as a pluricentric language. Under this account, Leitner uses an ‘epicentre’ metaphor to characterize the various regionally defined hubs in which emerging or emerged Englishes develop linguistic norms. These linguistic centres exert influence over the usage patterns of English in neighbouring areas. For example, in addition to American English (AmE) and British English (BrE), Leitner (1992) cites Indian English (IndE) and Singapore English (SingE) as two English epicentres from which linguistic behaviour diffuses to neighbouring areas, such as Pakistan and Sri Lanka or Malaysia and the Philippines. The results of the present study are consistent with this theory and also extend the selection of the regional epicentres that was initially advocated by Leitner. For example, we argue with the same logic that the West and East African countries that were part of the present study, although are more sparsely distributed in geographic space as compared to the Southeast Asian countries, show the hallmark of pluricentric norm-developed Englishes.

Our results are also compatible with the Dynamic Model of New Englishes, postulated by Schneider (2003a, 2007, 2010). This account sketches a process of ‘identity rewriting’ of both the indigenous and colonizing groups in the land that they jointly inhabit. Under this theory, a particular linguistic feature in a variety of English represents the end product of a historical process of the gradual appropriation, transformation and eventual indiginization of post-colonial Englishes. Describing the nativization process of World Englishes, Schneider (2010: 273) notes that

> a gradual diminishing in the social distance between English-speaking settler populations and indigenous populations emerges, due to the recognized need to share territory and life resources, and this process is reflected in language use, the symbolic use of forms of English... Consequently, contact forms of English can be observed emerging. Indigenous speakers of English transfer their own pronunciation habits, lexical expressions and also patterns of sentence composition from their respective native languages into their way of speaking English.

Thus, it could be the case that the pluralization of mass nouns (and other nouns in the data set) constitutes one type of ‘lexical expression’ that emerges via the above process. Moreover, it is also worth noting that the diffusion of
‘contact forms’ in Schneider’s (2003a: 271) nativization process rests on the assumption that the process of mutual accommodation between neighbouring linguistic communities also affects language change (Thomason and Kaufman 2001; Trudgill 1986). Therefore, the linguistic convergence observed between geographically proximate countries in Study 2 may reflect a process of cooperation between speakers of neighbouring countries, who tailor each other’s language production to be maximally successful in their online interaction.

The recreation of geographical distance from language usage statistics supports the idea that similar lexical forms are used by speakers from proximate locales. Yet, we wish to point out that the exact stage of development at which one English variety influences another cannot be answered by a synchronic data set such as ours. While Schneider’s (2003a) identity rewriting hypothesis is not incompatible with our data, the most parsimonious account is Leitner’s (1992). Leitner’s hypothesis, as far as we understand, does not restrict the point at which one variety of English influences another, to the stage at which a variety of English has been nativized. Future work with diachronic corpus linguistic data will be needed to pinpoint the exact nature of geo-linguistic diffusion of English varieties.

An alternative line of reasoning, foregrounded by Hall et al. (2013), proposes that L1 influence may partially explain apparent variability in noun countability. Hall et al. compared the magnitude of pluralization of 25 mass nouns among domains with classifier substrates and those with non-classifier substrates. Classifier languages (e.g., Bangladesh, Hong Kong, China) mark the noun referent with an additional word or affix in order to ‘classify’ if the noun is being counted. Hall et al. (2013) found that the overall level of pluralization was greater among non-classifier languages (e.g., Belize, Kenya, Malta). Based on this finding, they inferred a potential role of cross-linguistic influence, although an informal analysis of cross-linguistic transfer (whether or not pluralization of a noun in the dominant L1 correlates with its rate of pluralization in English) did not yield consistent patterns across OC varieties. Again, with the present data set and scope of analysis, we are unable to examine whether the acquisition of English plural morphology is subject to cross-linguistic influence or other factors of L2 learning, such as the process of overregularization and simplification within the mental lexicons of non-native English speakers.

It is clear that in order to reach a fuller understanding of the relationship between morpho-syntactic across varieties of English, future work will need distinguish historical from a-historical influences of language development. Cysouw (2013) recently offered a potential solution to an analogous problem in which he used a regression analysis to distinguish between genealogical from
geographical effects on language typology. Future approaches could fruitfully provide more answers to these problems by not defining speech communities at the country level. Perhaps data sources which contain more precise geo-spatial information for each language production event, such as Twitter, will be able to achieve a more granular snapshot of language contact between OC speech communities. Lastly, given the slight instability of the clustering outcome in Study 2, we urge future studies to model additional morpho-syntactic markers of non-native English in the same way. If convergence with the current findings, then this would provide support for the present results.

Moreover, the apparent difficulty in classifying varieties of English at the country-level points to some of the theoretical shortcomings of Kachru’s concentric circle model of World Englishes (for additional commentary see Crystal 1997; Mair 2013; Schneider 2007). One major theoretical concern of The Three Circles model is that it provides an outdated and over-simplified definition of ‘varieties’ of international English and does not reflect the true reality of World Englishes. This limitation is taken into account in more recent models of global Englishes. For example, Mair’s (2013) “World System of Englishes” incorporates new elements of a dynamic and emergent globalized linguistic community into Kachru’s model. Mair (2013) rightly points out that many varieties of English co-exist within a national boundary, which is a problem in The Three Circles model, which diagnoses native-speaker status based only the speaker’s country of provenance. In addition, Mair (2013) foregrounds the influence of digitally mediated communication on the variation of vernacular English, making the case that computer-mediated communication allows communication between speakers of non-varieties of English within and across territorial boundaries, and therefore must be considered as sub-system of World Englishes unto its own. Despite these theoretical limitations, we were not deterred by the coarse-grained IC versus OC dichotomy, as we believe that by partitioning countries in this way, we will still be able to capture a valuable bird’s-eye view of the gross differences in noun countability between native and non-native English speakers.

5 Conclusions

The present results demonstrate the utility of using a macro-level analytic framework in the task of retrieving a maximal set of nouns that are significantly more pluralized in the OC relative to the IC. However, identifying an underlying cause for the observed patterns of L2 English countability cannot be fully answered based purely on the statistical testing of word frequency counts.
In all, we argue that this paper lays a methodological groundwork for the macro-scale quantitative assessment of one morphological feature of non-native varieties of English. Supplementing the findings of this coarse-grained approach with a qualitative analysis of the semantic and syntactic factors of noun pluralization will contribute an invaluable source of knowledge to the investigation of morpho-syntactic behaviour in varieties of English.

Moreover, we argue that the motivation of this research goes beyond the provision of a statistically and quantitatively robust survey of noun countability in World Englishes. A study of the variability in morpho-syntactic behaviour of World Englishes, of which the pluralization of mass nouns is just one feature, also provides concrete data which can support the well-established challenge against ‘monolithic’ notions of the English language (Hall 2013, 2014; Kachru 1985; McKay and Bokhorst-Heng 2008; Makoni and Pennycook 2007). A monolithic conception of language assumes that there is a normed and idealized target variety of English and deviance from the target variety is deemed ‘erroneous’ or ‘deficient’. Applied linguists such as Hall (2014) and McKay and Bokhorst-Heng (2008) argue against the judgement of non-native speaker English against the so-called ‘standard’ native English speaker norms, an attitude still taken for granted by English Language Teaching (ELT) professionals and language researchers. With particular reference to noun countability, Hall et al. (2013) insist that upholding native-speaker norms in ELT practice is especially trivial, given that the presence or absence of plural morphology applied grammatically mass nouns serves little communicatively informative function. The current study was conducted in this spirit of English ‘plurilism’, which challenges the inherent notion of supremacy associated with native speaker Englishes.

In sum, the data presented here reflects the richly plurilithic and complex reality of World Englishes (Leitner 1992; Pennycook 2007). Our research supports the notion that the count–mass distinction, which is often conceived as a binary grammatical feature, is unworkable within a theory of linguistic representation. This conclusion is drawn from usage-based evidence which points towards a heterogeneous pattern of noun countability between Inner and Outer varieties of English. We found that mass nouns, defined as such from the perspective of native English norms, are most likely to more frequently pluralized by non-native speakers of English compared to native speakers of English. We also found that the relationship among Outer Circle countries with respect to the rate of occurrence of plural forms is influenced, in the broadest sense, by geography, and hence represents a non-random pattern. Given the pattern of our results, we propose that a fixed and binary count–mass distinction is not an essential component in a theory of English language structure. Instead, it is a phenomenon best viewed as a gradient that is also regionally dependent.
Acknowledgment: Thanks are due to Paweł Mandera for contributions to computational analysis in initial stages of this project. Thanks are also due to Christopher Hall, Anna Moro and Ivona Kucérová for their valuable comments on earlier drafts of this work, and to the attendees of the Conference of the Canadian Linguistics Association, Brock University, ON, Canada where this work was presented in May, 2014 and also to the attendees of the 9th International Conference on the Mental Lexicon, Niagara-on-the-lake, ON, Canada, September 2014. Thanks are also due to the students of the Department of Languages and Linguistics at York St. John University, UK, for their valuable comments on the early stages of this work.

Funding: This work was supported by the Ontario Trillium Award and a Graduate fellowship awarded by the Lewis & Ruth Sherman Centre for Digital Scholarship to the first author. The second author’s contribution was supported in part by the SSHRC Insight Development grant 430-2012-0488, the NSERC Discovery grant 402395–2012, the Early Researcher Award from the Ontario Research Fund, and the NIH R01 HD 073288 (PI Julie A. Van Dyke).

Appendix A

As stated in Statistical analyses, the main advantage of the LORIDP measure is that it shrinks the frequency of the words under comparison to a prior frequency in a large background corpus. This enables the researcher to detect differences between words that are very frequent. In order to verify that LORIDP does indeed correct for high-frequency items and also to support our implementation of the chosen measure, we compared the LORIDP measure to two competing measures that have also been used to discover the overpresentation of a word in one corpus as compared to another. For the set of 6,227 nouns that we identified in the data collection procedure, we computed the Damerau (1993) ratio of relative frequencies. This measure estimates the difference between the frequency of the plural form of noun \( w \) in the Outer Circle \( i \), and Inner Circle \( j \). The relative frequency ratio \( r \) is computed as,

\[
r = \frac{\frac{\hat{y}_i}{n_i}}{\frac{\hat{y}_j}{n_j}},
\]

where \( n_i \) is the total number of noun tokens in the Inner Circle subdivision of the GloWbE corpus \( i \), \( n_j \) is the total number of noun tokens in the Outer Circle subdivision of the GloWbE corpus \( j \).
We also computed a simple difference coefficient used by Johansson and Hofland (1982) and Leech and Fallon (1992), which is computed as,

$$\frac{y^i_w - y^j_w}{y^i_w + y^j_w}.$$

To test whether LORIDP affords an advantage over other difference coefficient measures, we separately correlated total word frequency of each word in our data set with each difference coefficient of the over- or underrepresentation of the plural form of the word in Outer Circle Englishes. As the LORIDP measure shrinks word frequency, it is therefore not expected to correlate as strongly with word frequency when compared to the other measures. If the total frequency does not correlate well with LORIDP, then this validates LORIDP as a measure that estimates the difference in pluralization across the Inner and Outer Circle while controlling for the effect of total frequency. As is reported in Table 3, the LORIDP measure yields the weakest correlation with word frequency. We take this observation as support for the adoption of the LORIDP measure in the present investigation.

Table 3: Spearman’s $\rho$ for the correlation between each difference metric and the summed frequency of the plural and singular form of nouns in GloWbE.

<table>
<thead>
<tr>
<th>Difference coefficient metric</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Damerau’s $r$</td>
<td>-0.17</td>
</tr>
<tr>
<td>Johansson’s difference coefficient</td>
<td>-0.19</td>
</tr>
<tr>
<td>Log-odds ratio informative Dirichlet prior (LORIDP)</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

References


