

# Outer and expanding circle Englishes

## The competing roles of norm orientation and proficiency levels\*

Alison Edwards and Samantha Laporte

University of Cambridge / Université catholique de Louvain

The classification of English as a native (ENL), second (ESL) and foreign (EFL) language is traditionally mapped onto Kachru's (1985) Inner, Outer and Expanding circles, respectively. This paper addresses the divide upheld between these different varietal types. We explore the preposition *into* using comparable corpora for all three varietal types: the International Corpus of English (ICE) for Inner and Outer Circle varieties, and a comparable Corpus of Dutch English to represent the Expanding Circle. Our results show that the least institutionalised varieties (Hong Kong and Dutch English) are the most dissimilar to the ENL varieties, and the most institutionalised variety (Singapore English) is the most similar. We also compare our results for the Corpus of Dutch English to the Dutch component of the International Corpus of Learner English. While the latter patterns with other learner varieties, the Dutch English corpus patterns with ESL varieties, suggesting that "Expanding Circle" and "EFL" are not synonymous.

**Keywords:** EFL, ESL, ENL, Outer Circle, Expanding Circle, norm orientation, prepositions, Netherlands

### 1. Background

With this paper, we aim to reassess some of the assertions made in the literature with respect to the cline said to hold between, and the characteristics attributed

---

\* We are grateful to Sandra Deshors for helpful discussion at the 34th ICAME conference in Santiago de Compostela, and to Gaëtanelle Gilquin, Sylviane Granger, Marianne Hundt and two anonymous reviewers for their constructive comments on earlier versions of the manuscript. Any remaining errors are our own.

to, varieties of English as a native (ENL), second (ESL) and foreign (EFL) language. These varietal types are typically loosely mapped onto Kachru's (1985) Inner, Outer and Expanding circles, respectively. ESL varieties are those that have emerged in postcolonial contexts and are considered to be 'norm developing'; that is, their users are seen as agentively shaping the language for their own sociocultural ends. In contrast, EFL varieties are regarded as 'norm dependent'; they are considered 'learner' or 'performance' varieties whose speakers continue to look to ENL varieties for their target norms. It has recently been claimed that as a result of these divergent norm orientations — ESL as endonormative, EFL as exonormative — ESL varieties will show systematically different structural properties compared to ENL, whereas EFL varieties will behave in a relatively more native-like fashion (Van Rooy 2006: 37; Hundt and Vogel 2011: 157, see below). However, it remains unclear how proficiency levels might play out in this; an EFL variety, for example, may be exonormative in orientation, but due to low overall proficiency levels highly dissimilar from ENL in its actual performance. Further, any account that hinges on purported norm orientation should also cater for the fact that ENL norms are themselves by no means homogeneous.

The claim that their respective norm orientations will cause ESL varieties to be less and EFL varieties more similar to ENL varieties is illustrated by two studies of progressive aspect marking. First, Hundt and Vogel (2011) compared progressive marking in student writing from Fiji, Kenya, Malaysia, the Philippines and Singapore (ESL), Germany, Finland and Sweden (EFL), and Great Britain, Ireland and New Zealand (ENL). In terms of overall frequencies they found no clear grouping along the lines of varietal type. Qualitatively, however, they observed that progressives are used in a 'creative way' in the ESL varieties, showing 'stretched' tolerance towards new aspectual uses and combinations of the progressive with stative verbs. In contrast, they stated "this is not the case in learner English", asserting that EFL learners are "more likely to overuse the prototype of the construction and less likely to 'stretch' the progressive to new contexts" (Hundt and Vogel 2011: 160). In other words, their qualitative analyses revealed the EFL varieties to be more exonormative than the 'norm developing' ESL varieties. Likewise, Van Rooy (2006) investigated ESL and EFL varieties using data from, respectively, the Tswana Learner English Corpus (TLE)<sup>1</sup> and ICLE Germany. He found that German learners made greater use of prototypical progressive constructions denoting action in progress, while the Tswana speakers used the progressive "in very different ways", developing a different constructional prototype expressing "a kind

---

1. It is worth noting that Van Rooy (2006) considers Tswana L1s as ESL users, whereas Gilquin and Granger (2011; see Section 2), consider them to be EFL learners. In discussing these studies we use the classifications upheld by the respective authors.

In this paper we further explore the cline in (2). Are the results on which it is based (Hundt and Vogel 2011; Mukherjee and Gries 2009; Van Rooy 2006) valid for the structural features investigated in those studies only, or do they hold for other features too? Is norm orientation really the decisive factor in determining a variety's location on the cline, or do additional factors, such as proficiency levels, also play a role? And can we really differentiate so clearly between the tripartite categories of ENL, ESL and EFL? In recent years, numerous researchers have suggested that the strict dividing lines between varietal types be reconceptualised in favour of a continuum (e.g. Biewer 2011; Buschfeld 2011; Gilquin and Granger 2011) and that, as a corollary, comparative analyses should consider varieties of all types (see Buschfeld 2011; Davydova 2012; Mukherjee and Hundt 2011; Nesselhauf 2009). In investigating the cline in (2), therefore, we address this by incorporating data from all three of Kachru's (1985) circles.

## 2. Study design and methods

In this paper, we explore the proposed cline in (2) for another linguistic phenomenon: the preposition *into*. Differences in the distribution of prepositions have been found even among ENL varieties (see Peters 1998); for example, Hofland and Johansson (1982: 505) found *into* to be slightly more frequent in AmE compared to BrE. In New Englishes, prepositions are seen as a “mutating species”, so often do they give rise to innovations (Gilquin and Granger 2011: 60). Different ESL varieties — East African English (Mwangi 2003), Pakistani English (Mahboob 2008) and Indian English (Mukherjee 2009), to name just a few — are reportedly characterised by omission, insertion or substitution of prepositions. In EFL, “prepositions are often considered as the *bête noire* of both teachers and learners, being impossible to teach and impossible to learn” (Gilquin and Granger 2011: 60). Interestingly, *into* in particular has been identified in various innovative usages and/or ‘errors’ in studies of ESL and EFL varieties (e.g. Gilquin and Granger 2011; Nesselhauf 2009). One apparent reason is its semantic link with *in*, which could give rise to confusion exacerbated if the speaker’s L1 does not distinguish between prepositions of location and direction (see e.g. Mwangi 2003 on East African English).

Our starting point is a study by Gilquin and Granger (2011), who investigated the patterning of *into* in the International Corpus of Learner English components for Dutch (ICLE-DU), French (ICLE-FR), Spanish (ICLE-SP) and Tswana (ICLE-TSW) learners, compared to a reference corpus of BrE editorials (the Multilingual Editorials Corpus, MULT-ED). Their results reflected the different levels of exposure to English and focus on form during instruction in the respective countries: ICLE-DU was typically the most similar to BrE, followed by ICLE-FR and then ICLE-SP. These findings suggest a cline among EFL varieties whereby higher proficiency levels equate to more native-like performance.<sup>2</sup> This represents an interesting counterpoint to Mukherjee and Gries (2009) as discussed above, whose cline among ESL varieties was largely attributed to matters of norm orientation. It is our hope that integrating different varietal types in the present study will help us to disentangle these apparently competing factors of norm orientation and proficiency level.

As noted above, the data used in the present study are therefore drawn from corpora representing all three of Kachru’s (1985) circles. Representing the Inner and Outer circles are data from the written components of the International

---

2. A rating of a sample of 20 ICLE essays per L1 background according to the Common European Framework of Reference suggests a proficiency cline that runs parallel to the amount of (i) exposure to English and (ii) focus on form in instruction for ICLE-SP, ICLE-FR and ICLE-DU (Granger et al. 2009: 11–12).

Corpus of English (ICE; Greenbaum 1991) for Britain (ICE-GB), the United States (ICE-USA), Hong Kong (ICE-HK), India (ICE-IND) and Singapore (ICE-SIN).<sup>3</sup> The Expanding Circle is represented by a new Corpus of Dutch English (NL) (Edwards 2011). Table 1 provides an overview of the corpora used and their respective sizes.

**Table 1.** Overview of the corpora and their word counts\*

| Type | Corpus  | Word count |
|------|---------|------------|
| ENL  | ICE-GB  | 422,622    |
|      | ICE-USA | 419,183    |
|      | ICE-SIN | 399,350    |
| ESL  | ICE-IND | 411,491    |
|      | ICE-HK  | 491,658    |
| EFL  | NL      | 401,199    |

\* See Edwards (2011) for the method used to obtain the word counts.

## 2.1 The Corpus of Dutch English (NL)

Over the years, various authors (e.g. Ammon and McConnell 2002: 99; Booij 2001; Jenkins 2009: 16–17; Kirkpatrick 2007: 165; McArthur 1998: 54; Mesthrie and Bhatt 2008: 211) have noted in passing that the Netherlands may be transitioning from EFL to ESL status. In the latest Eurobarometer report (European Commission 2012), 90 per cent of Dutch respondents reported being able to hold a conversation in English. In the English Proficiency Index, a ranking of 54 countries in which English is not the national language, the Netherlands was classed in the “very high proficiency” group, ranking a close third behind Sweden and Denmark (Education First 2012: 4). Moreover, various societal domains in the Netherlands, such as business, advertising and the media, operate under the “assumption of bilingualism” of the population, which is suggestive of ESL; this both allows for, and increasingly gives rise to, expressive and creative uses (Edwards 2011).

However, as English is not passed on naturalistically from parent to child as in the Inner Circle, and given the absence of a colonial history in the Outer Circle

3. The written components of the ICE corpora were used in order to be comparable with the Corpus of Dutch English, which is based on the ICE design but currently only consists of a written component. It is hoped that a spoken component will be added in due course (Edwards 2011). While emergent changes are said to be first traceable in spoken language, any incipient norms observed in writing could provide even stronger evidence of endonormative developments. However, it should be noted that the analyses, results and implications discussed in this paper refer only to written language.

sense, the Netherlands is traditionally considered an Expanding Circle country. As a result, it does not fall within the scope of the ICE project, which expressly includes only those countries where English is “the first language or a second official language”.<sup>4</sup> The compilation of the NL corpus ties in with recent calls to develop corpus data for World Englishes (WEs) research that better reflect the changed sociolinguistic realities in Expanding Circle countries; as Laitinen (2011) writes, “there is a need to reconsider the existing stock of data used in the field [...] It is essential that new corpora, which match the global spread of the language, are developed”.

The NL corpus was therefore built at the University of Cambridge, with 200 texts divided over eight different genres, totalling approximately 400,000 words (Table 2) (Edwards  2011). This not only makes it readily comparable with the ICE corpora, but also allows for comparisons with an Expanding Circle corpus that is not restricted to learner writing only. The texts were collected over six months in 2011 and, like the second ‘wave’ of ICE corpora, date from 2005 onwards. Although the time lag compared to the first wave of ICE corpora from the early 1990s is not desirable, it was considered unavoidable at this time.<sup>5</sup> Contributors were L1 Dutch speakers who were born and raised in the Netherlands or had moved there as infants, and who had not lived outside the Netherlands for more than 10 years or over half their lifetimes (see Holmes 1996).<sup>6</sup> All had received at least five years of instruction in English as a foreign language. Around 90 per cent had obtained an undergraduate degree or higher, almost 40 per cent of whom had followed all or part of their higher education in English. Personal information — age, sex, home town/city, time spent abroad, parents’ languages, language(s) of instruction during schooling and higher education (if applicable), education level and occupation — was collected by means of a questionnaire and included as metadata in the text headers. As with the ICE corpora, a cline of bilingualism can be assumed that is associated with the communicative uses to which English is put by the students, businesspeople, journalists, academics and others who contributed to the corpus.

---

4. <<http://ice-corpora.net/ice/joinice.htm>> (accessed March 16, 2014)

5. The first ‘wave’ of ICE components dates from the early 1990s. This was followed by a second wave, dubbed *ICE Age 2*, with texts dating from 2000 or 2005 (see the articles in the special issue of *ICAME journal ICE Age 2: ICE corpora of New Englishes in the making*, volume 34). The texts in the NL corpus are therefore of the same ‘era’ as the second wave of ICE corpora.

6. Contributors were also asked when and where they had spent time abroad, and this information is included in the metadata in the corpus text headers. Approximately 8 per cent had spent one to three years and 7 per cent more than three years in an English-speaking country (Edwards 

Table 2. Corpus of Dutch English

| Category                  | Texts | Words   |
|---------------------------|-------|---------|
| Correspondence            |       |         |
| Social                    | 15    | 29,862  |
| Business                  | 15    | 28,786  |
| Student writing           |       |         |
| Untimed essays            | 10    | 20,283  |
| Examination essays*       | 10    | 20,143  |
| Academic writing          |       |         |
| Humanities                | 10    | 20,967  |
| Social sciences           | 10    | 20,145  |
| Natural sciences          | 10    | 18,697  |
| Technology                | 10    | 19,846  |
| Popular writing           |       |         |
| Humanities                | 10    | 20,296  |
| Social sciences           | 10    | 19,712  |
| Natural sciences          | 10    | 20,478  |
| Technology                | 10    | 20,379  |
| Reportage                 |       |         |
| Press news reports        | 20    | 40,401  |
| Persuasive writing        |       |         |
| Press editorials          | 10    | 20,670  |
| Instructional writing     |       |         |
| Administrative/regulatory | 10    | 19,218  |
| Skills/hobbies            | 10    | 20,411  |
| Creative writing          | 20    | 40,905  |
| Total                     | 200   | 401,199 |

\* The examination essays used in the corpus analysis reported in this study were drawn under licence from the ICLE-DU. They remain the property of the ICLE project and are not officially part of the NL corpus.

Only one study using the NL corpus has been conducted to date, focusing on the progressive aspect compared to several ICE corpora (Edwards 2000). Dutch English showed a similar extension of progressive marking to stative verbs and contexts of habitual activity or general validity as in the ESL corpora. The findings therefore revealed no strict divide between the ESL corpora on the one hand and NL on the other; indeed, ICE-SIN consistently performed in the most exonormative manner (see also Hundt and Vogel 2011) and ICE-IND the least, with NL in between.

These recent findings notwithstanding, in the present study on the preposition *into* we hypothesised a cline derived from the results in Hundt and Vogel (2011), Van Rooy (2006) and Mukherjee and Gries (2009), as shown in (3) below. NL, as

an Expanding Circle variety, was hypothesised to be the most exonormatively oriented. In line with Mukherjee and Gries (2009), the cline among the ESL varieties was hypothesised to reflect their advancement in Schneider’s (2003, 2007) model, with ICE-HK (phase 2/3) being the most exonormatively oriented and thus the most similar of the ESL varieties to ENL, and ICE-SIN (phase 4) being the most endonormatively institutionalised and thus the most dissimilar to ENL.

(3) USA/GB – NL – HK – IND – SIN

2.2 Data extraction and coding

In line with Gilquin and Granger (2011), we explored six areas of interest: the overall frequency of *into*, its syntactic and semantic distribution, lexical variation of the verbal structures used with *into*, phraseological uses and nonstandard uses of *into*. First, concordances with *into* were extracted from each corpus using AntConc (Laurence Anthony, version 3.2.4), and the results exported to Microsoft Excel 2010. Occurrences of *into* in quotes and book titles were excluded to arrive at the overall frequency counts. For the purpose of comparison all concordances were then manually coded in line with the categorisations in Gilquin and Granger (2011), as described in the subsections below.

2.2.1 Syntactic distribution

Four categories were established into which each occurrence of *into* falls: ‘noun + *into*’, ‘transitive verb + *into*’, ‘intransitive verb + *into*’, or ‘other’. Table 3 illustrates each of the structures using examples from the corpora.

Table 3. Syntactic classification of *into*

| Structure                          | Example                                                                                                                                    |
|------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| N + <i>into</i>                    | [T]his would not in itself justify the huge explosion in <i>research into</i> artificial neural nets which is seen today. (ICE-GB W2F-017) |
| V <sub>trans</sub> + <i>into</i>   | The cellphone features a small generator that is able to <i>convert</i> movement into energy smart thinking. (NL W2B-033)                  |
| V <sub>intrans</sub> + <i>into</i> | Days <i>melt</i> into months and months into years. (ICE-USA W2B-011)                                                                      |
| Other                              | Then up the steps, through the double doors and into a wide, brightly-lit corridor. (ICE-GB W1B-003)                                       |

2.2.2 Lexical variation

The focus here is on the verbal structures already coded above as ‘transitive/intransitive verb + *into*’, as these represent nearly 90 per cent of the uses of *into* in all corpora (see also Gilquin and Granger 2011). The different verb lemmas were

retrieved and recorded. Phrasal verbs were counted as separate verb lemmas (e.g. *that habit will carry on into your dreams*, USA W2D-013) and were distinguished from free combinations of 'verb + adverb', which were recorded under the verb lemma only (e.g. *a line of slim-hipped men in evening wear filed past into the hall*, GB W1B-012).<sup>7</sup>

### 2.2.3 Semantic distribution

Each occurrence of *into* was classified into one of eight categories, which Gilquin and Granger (2011) established on the basis of their ICLE data<sup>8</sup> in an approach comparable to that in De Cock and Granger (2004). The present study confirms the robustness of their categories; we used an iterative approach, starting from the corpus data, confirming intuitions about phrasal verbs and (semi-)fixed expressions with dictionaries,<sup>9</sup> and continually revising the classification of occurrences to maximise consistency.<sup>10</sup> Table 4 provides a brief explanation of and examples from the corpora for each semantic category.

7. Following Quirk et al. (1985: 1152–5), if (1) the verb could easily be substituted by other verbs, and/or (2) the modifying adverb *right* could be inserted between the verb and the particle and/or (3) the adverb could be placed before the (intransitive) verb, it was considered a combination of 'verb + adverb' and only the verb was recorded.

8. Gilquin and Granger (2011: 64) also used the Oxford Advanced Learner's Dictionary and the Macmillan English Dictionary for Advanced Learners, Second Edition as references.

9. (Semi-)fixed expressions were classed as such if they were listed under *idioms* in the Oxford Advanced Learner's Dictionary (oald8.oxfordlearnersdictionaries.com) or *phrases* in the Oxford Dictionary (www.oxforddictionaries.com).

10. Though such a classification necessarily involves a degree of subjectivity, measures were taken to ensure consistency as far as possible. First, each of the two authors classified the same 100 occurrences individually. Comparison revealed an inter-rater agreement of 74 per cent, which increased to 96 per cent after discussion. We then sought the advice of Gaëtanelle Gilquin to clarify how the remaining 4 per cent of ambiguous cases had been classified in Gilquin and Granger (2011). Subsequently, the first author classified all occurrences of *into* in each of the six corpora. To maintain consistency, 'prototypical' corpus examples of each semantic class were saved in a spreadsheet which was used as a reference and frequently updated throughout the process of classification. Next, the second author conducted a post-hoc check of 100 further random occurrences. Inter-rater agreement was 88 per cent, which increased to full agreement after discussion. Finally, some persistently difficult cases were re-checked by the first author.

Table 4. Semantic classification of *into*

| Sense/use                | Explanation                                                                                                    | Corpus example                                                                                               |
|--------------------------|----------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Movement                 | Verbs of movement (e.g. <i>come</i> , <i>go</i> , <i>step</i> ) used with a literal meaning                    | A weekend later he <i>crept</i> into her bedroom and won his bet. (ICE-GB W2F-017)                           |
| Abstract movement        | Verbs of movement (e.g. <i>come</i> , <i>go</i> , <i>step</i> ) used with a metaphorical meaning               | Ludwig <i>came</i> into the ceramics world through the back door. (ICE-USA W2B-004)                          |
| Transformation           | Verbs denoting some form of transformation, e.g. <i>change</i> , <i>develop</i> , <i>translate</i>             | In Holland [...] we <i>turn</i> anything with a little historical significance into a landmark. (NL W2B-006) |
| Causation                | Causal verbs, e.g. <i>cajole</i> , <i>force</i> , <i>trigger</i>                                               | Drivers are <i>blackmailed</i> into paying 100 yuan each, before being allowed to pass. (ICE-HK W2B-015)     |
| Division                 | Verbs denoting some form of division, e.g. <i>break</i> , <i>classify</i> , <i>split</i>                       | Twentieth-century women's novels can be roughly <i>divided</i> into two periods. (ICE-GB W2B-009)            |
| Other meanings           | All senses that cannot be classified otherwise, e.g. <i>incorporate into</i> , <i>gain insight into</i>        | This was <i>continued</i> into the Ming-Ching period. (ICE-SIN W1A-017)                                      |
| Phrasal verbs            | Non-composition prepositional verbs, e.g. <i>break into</i> , <i>feed into</i> , <i>look into</i>              | You are requested to <i>look</i> into the matter. (ICE-IND W1B-027)                                          |
| (Semi-)fixed expressions | Idioms and phrases, e.g. <i>fall into a trap</i> , <i>go into overdrive</i> , <i>play into someone's hands</i> | Body mass, height and gender were <i>taken</i> into account as known predictors. (NL W2A-022)                |

#### 2.2.4 Phraseological uses

This analysis zooms in on the (semi-)fixed expressions identified in the semantic classification above. We examine the most frequent of these and, given the low frequency counts, support this with qualitative discussion.

#### 2.2.5 Nonstandard uses

Occurrences of *into* were classified as either standard or nonstandard in the native as well as the non-native corpora, in recognition of the fact that native speakers, too, do not always use English in strictly conventional ways. The general principle in determining standardness was whether the relevant collocation appeared in a dictionary and/or was the conventionally favoured form both in terms of

native-speaker intuition and in domain-specific Google searches.<sup>11</sup> As in Gilquin and Granger (2011), this classification relies on the judgement of one native speaker only. Thus, while steps were taken to ensure consistency and objectivity as far as possible,<sup>12</sup> the results should be seen as indicative.

### 3. Results

#### 3.1 Overall frequencies

Table 5 shows the raw and normalised frequencies of the preposition *into* across all corpora, with figures per 100,000 words ranging from a low of 128 in NL up to 159 in ICE-SIN (see also Figure 1, with error bars representing 5 per cent confidence intervals).<sup>13</sup> A Poisson regression modelling process shows that in the most significant and economic model (LRT = 32.537,  $df = 1$ ,  $p < 0.001$ ), the only significant contrast is that between ICE-USA, ICE-SIN and ICE-GB on the one hand, and ICE-IND, ICE-HK and NL on the other.<sup>14</sup> In other words — and contrary to our hypothesised cline in (3) — the *least* institutionalised varieties, ICE-IND, ICE-HK and NL, differ significantly from the ENL varieties, while the *most* institutionalised variety, ICE-SIN, does not. This is in line with Gilquin and Granger's (2011) finding of general underuse of *into*, with the least institutionalised varieties in their data also showing the greatest underuse. However, their relative figures (ranging from the low 70s to the low 100s) for various ICLE corpora were

11. This means that less common uses of *into*, though not necessarily wrong, were classed as nonstandard, e.g. *migrate into* (rather than *to*), *analysis into* (rather than *of*), *information into* (rather than *about*, *on*).

12. The first author coded each occurrence of *into* as either standard or nonstandard in two rounds approximately four months apart. An intra-rater disagreement rate of 3.7 per cent (135 out of a total of 3689 concordances) was found between rounds. The first round of classification was done per corpus; thus, the rater was aware which corpus each occurrence of *into* came from. As it was felt that this might bias the results, in the second round the concordances were randomised so that the rater was blinded as to which corpus they came from, and indeed this saw the nonstandard proportions in ICE-USA and ICE-GB corpora approximately double (from around 1.2 per cent to 2.6 per cent in each), though the rank order across varieties remained the same.

13. Relative frequencies are reported per 100,000 words to ensure that the figures are readily comparable to those of Gilquin and Granger (2011).

14. The functions used for this procedure were *glm* to fit the regression models, and *anova* to compare models when conflating levels of the variety variable. Both functions are part of R's stats package. We thank an anonymous reviewer for suggesting this procedure.

Table 5. Raw and normalised frequencies of *into*

| Corpus  | Into: raw | Into: per 100,000 words |
|---------|-----------|-------------------------|
| ICE-GB  | 665       | 157.4                   |
| ICE-USA | 665       | 158.6                   |
| ICE-SIN | 636       | 159.3                   |
| ICE-IND | 565       | 137.3                   |
| ICE-HK  | 644       | 131.0                   |
| NL      | 512       | 127.6                   |

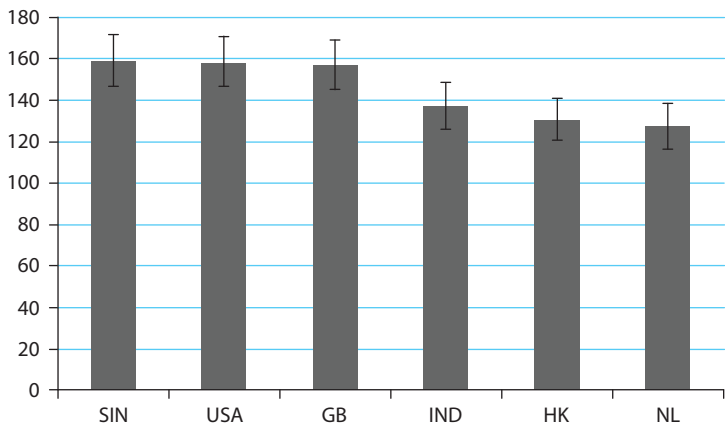


Figure 1. Relative frequency of *into* per 100,000 words across all corpora

considerably lower overall than those for the present corpora. Consider, for example, Gilquin and Granger’s (2011) relative frequency of 102 for ICLE-DU (consisting of undergraduate essays), compared to the present figure of approximately 128 for the NL corpus. This seems to suggest that the major factor at play in the use of the preposition *into* is a proficiency effect: the lower the proficiency level, the greater the underuse of *into*.

3.2 Syntactic and semantic distribution

Following Gilquin and Granger (2011), we investigated the distribution of *into* across different syntactic structures and semantic classes. As demonstrated by Mukherjee and Hoffmann (2006:248), nativisation manifests itself not only in qualitative innovations but also through “shifts in frequency” of certain phenomena. Therefore, echoing the Behavioral Profile approach outlined in Gries (2010), we conducted hierarchical cluster analyses (HCA) to identify any quantitative preferences in syntactic and semantic patterning across varieties. HCA is an exploratory statistical technique used to measure similarity and dissimilarity between groups, or in this case, corpora, providing a ‘bird’s eye perspective’ of the entire dataset in

the form of a dendrogram.<sup>15</sup> Previous applications of HCA in variationist studies can be found in e.g. Mukherjee and Gries (2009) and Werner (2011). All dendrograms in the paper were produced in R using the function `pvclust`; for related applications using this function, see also Gries (2010).<sup>16</sup>

### 3.2.1 Syntactic distribution

Table 6 reports the relative frequencies (per 100,000 words) and the percentage distributions of each syntactic structure. The overall rank order of the structures is the same across all corpora, with the transitive verbal structure predominating, followed by the intransitive verbal and then the nominal construction. Based on this frequency data, the dendrogram in Figure 2 shows how the varieties relate to one another in terms of syntactic distribution (see Appendix 1 for the input data). The box indicates that ICE-GB, ICE-HK and ICE-USA, with very similar distributions, differ significantly from ICE-SIN, NL and ICE-IND. The cluster formed by NL and ICE-IND can be explained by the fact that while the transitive verbal structure is significantly more frequent than the intransitive structure in all other corpora,<sup>17</sup> NL and ICE-IND have a much more even distribution of these two patterns.

The dendrogram as a whole cannot readily be mapped onto a cline according to the degree of institutionalisation of each variety. Hence, it neither confirms our hypothesised cline, nor does it reflect the proficiency effect noted in the previous section. It is noteworthy, however, that the clusters cut across varietal types. ICE-HK, the least institutionalised Outer Circle variety, clusters with the Inner Circle varieties, followed by ICE-SIN, the most institutionalised Outer Circle variety. On the other branch ICE-IND, an Outer Circle variety, clusters with NL, the Expanding Circle variety. We will return to this point later.

15. These are based on the distribution across the syntactic and semantic classes per corpus. They do not reflect relative frequencies of each syntactic/semantic category out of the total number of words in each corpus, but rather the relative frequencies of each category out of the number of *intos* per corpus (i.e. the relative frequency of each category relative to the other categories in that corpus). This prevents the dendrogram from merely reflecting the overall underuse of *into* noted for certain varieties in the previous section.

16. The function `pvclust` (available in the `pvclust` package for R (<http://cran.r-project.org/web/packages/pvclust/index.html>)), provides two *p*-values: A “bootstrap probability” (BP) value and an “approximately unbiased” (AU) value. The latter is computed on the basis of multiscale bootstrap resampling and is less subject to bias than the BP value (Suzuki and Shimodaira 2006). The significant clusters are therefore determined on the basis of the AU values ( $AU \geq 95$ , i.e.  $p < 0.05$ ).

17. For example, the values for ICE-USA, the corpus with the smallest difference between the transitive and intransitive structures, are  $\chi^2 = 13.66$ ,  $df = 1$ ,  $p < 0.001$ ,  $\phi_c = 0.101$ .

Table 6. Distribution of syntactic structures of *into* per corpus

| Corpus     | V <sub>trans</sub> + <i>into</i> | V <sub>intrans</sub> + <i>into</i> | N + <i>into</i> | Other | Total |
|------------|----------------------------------|------------------------------------|-----------------|-------|-------|
| ICE-GB     |                                  |                                    |                 |       |       |
| raw        | 352                              | 251                                | 72              | 10    | 665   |
| normalised | 78.6                             | 59.4                               | 17.0            | 2.4   | 157.4 |
| %          | 49.9                             | 37.7                               | 10.8            | 1.5   | 100.0 |
| ICE-USA    |                                  |                                    |                 |       |       |
| raw        | 330                              | 263                                | 65              | 7     | 665   |
| normalised | 78.5                             | 62.7                               | 15.5            | 1.7   | 158.4 |
| %          | 49.5                             | 39.5                               | 9.8             | 1.1   | 100.0 |
| ICE-SIN    |                                  |                                    |                 |       |       |
| raw        | 340                              | 248                                | 46              | 2     | 636   |
| normalised | 85.1                             | 62.1                               | 11.5            | 0.5   | 159.3 |
| %          | 53.5                             | 39.0                               | 7.2             | 0.3   | 100.0 |
| ICE-IND    |                                  |                                    |                 |       |       |
| raw        | 263                              | 256                                | 42              | 4     | 565   |
| normalised | 64.0                             | 62.3                               | 10.2            | 1.0   | 137.5 |
| %          | 46.5                             | 45.3                               | 7.4             | 0.7   | 100.0 |
| ICE-HK     |                                  |                                    |                 |       |       |
| raw        | 327                              | 243                                | 69              | 5     | 644   |
| normalised | 66.5                             | 49.4                               | 14.0            | 1.0   | 131.0 |
| %          | 50.8                             | 37.7                               | 10.7            | 0.8   | 100.0 |
| NL         |                                  |                                    |                 |       |       |
| raw        | 230                              | 210                                | 64              | 8     | 512   |
| normalised | 57.3                             | 52.3                               | 16.0            | 2.0   | 127.6 |
| %          | 44.9                             | 41.0                               | 12.5            | 1.6   | 100.0 |

3.2.2 Semantic distribution

Table 7 shows the normalised frequencies and percentage distributions of *into* per semantic class for all six corpora, while Figure 3 presents the dendrogram calculated from the HCA of the semantic distribution (see Appendix 2 for the input data). The dendrogram clearly shows that the only variety that differs significantly from the others is NL. This can be explained by the frequency data in Table 7, which show that concrete movement is the predominant sense in all corpora except NL, which uses *into* in the movement sense significantly less than all other corpora ( $\chi^2=26.84$ ,  $df=5$ ,  $p<0.001$ ,  $\phi_c=0.085$ ). This is in contrast to Gilquin and Granger (2011: 64), who found for their data that

the prototypical sense of concrete movement is never predominant. Instead, the most frequent sense in the [BrE] reference corpus is abstract movement [...] whereas in the four ICLE components it is the (semi-)fixed expressions that are most common.

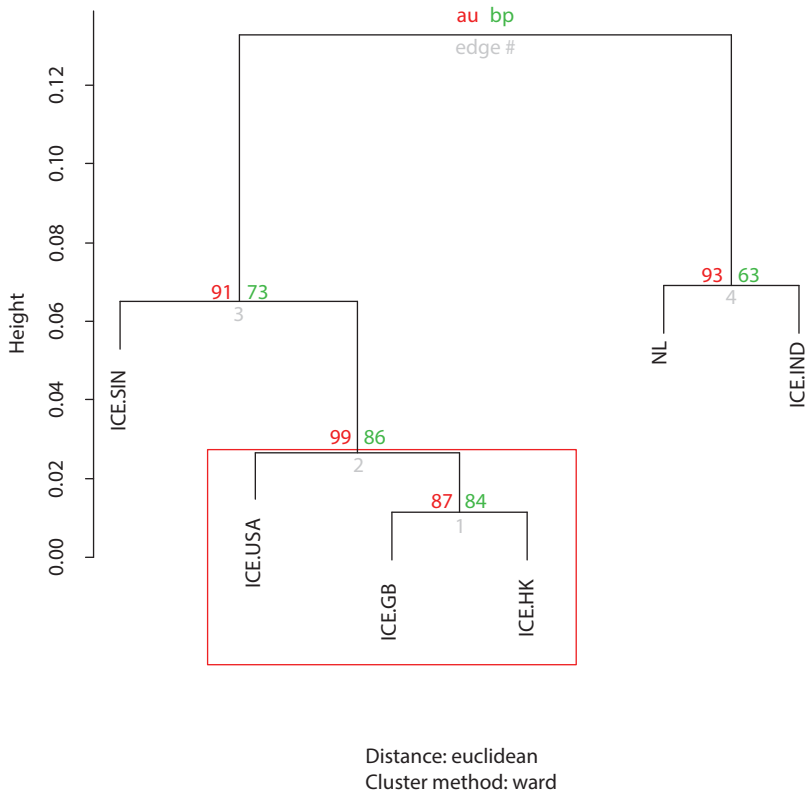


Figure 2. Hierarchical cluster analysis of syntactic distribution of *into*

In the present data, none of the four non-native corpora use (semi-)fixed expressions significantly more than the ENL corpora. This could be linked to proficiency levels; the learners in Gilquin and Granger (2011) showed more systematically divergent results compared to their BrE reference corpus than do the present educated non-native speaker (NNS) corpora compared to ICE-GB and ICE-USA. However, while the difference is not significant, NL and ICE-HK do seem to show a trend towards more (semi-)fixed expressions, which is explored further in Section 3.4.

Like the dendrogram for the syntactic distribution in Figure 2, the dendrogram representing the semantic distribution in Figure 3 shows little connection between the clustering of the varieties and their degree of institutionalisation, with the exception of NL — the least institutionalised variety — which is isolated. This heterogeneity may be attributable to the size of the present corpora. Much larger corpora may yield more reliable results when it comes to analyses of semantic aspect. It is worth pointing out, in addition, that the two native varieties do not cluster closest together (although the difference between them is not statistically

Table 7. Distribution of semantic classes of *into* per corpus

| Corpus            | movement                                                                          | abstract<br>movement | transformation | causation | division | other meanings | phrasal verbs | (semi-)fixed<br>expressions | total |
|-------------------|-----------------------------------------------------------------------------------|----------------------|----------------|-----------|----------|----------------|---------------|-----------------------------|-------|
| ICE-GB            |  |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        |                                                                                   | 77                   | 94             | 31        | 54       | 91             | 61            | 54                          | 665   |
| <i>normalised</i> | 48.0                                                                              | 18.2                 | 22.2           | 7.3       | 12.8     | 21.5           | 14.4          | 12.8                        | 157.4 |
| %                 | 30.5                                                                              | 11.6                 | 14.1           | 4.7       | 8.1      | 13.7           | 9.2           | 8.1                         | 100.0 |
| ICE-USA           |                                                                                   |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        | 167                                                                               | 133                  | 110            | 20        | 36       | 98             | 55            | 46                          | 665   |
| <i>normalised</i> | 40.0                                                                              | 31.6                 | 26.2           | 4.8       | 8.6      | 23.3           | 13.1          | 10.9                        | 158.4 |
| %                 | 25.2                                                                              | 20.0                 | 16.5           | 3.0       | 5.4      | 14.7           | 8.3           | 6.9                         | 100.0 |
| ICE-SIN           |                                                                                   |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        | 179                                                                               | 90                   | 119            | 14        | 46       | 82             | 59            | 47                          | 636   |
| <i>normalised</i> | 44.8                                                                              | 22.5                 | 29.8           | 3.4       | 11.5     | 20.5           | 14.8          | 11.8                        | 159.3 |
| %                 | 28.1                                                                              | 14.2                 | 18.7           | 2.2       | 7.2      | 12.9           | 9.3           | 7.4                         | 100.0 |
| ICE-IND           |                                                                                   |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        | 157                                                                               | 87                   | 112            | 17        | 48       | 46             | 55            | 43                          | 565   |
| <i>normalised</i> | 38.2                                                                              | 21.2                 | 27.3           | 4.1       | 11.7     | 11.2           | 13.4          | 10.5                        | 137.5 |
| %                 | 27.8                                                                              | 15.4                 | 19.8           | 3.0       | 8.5      | 8.1            | 9.7           | 7.6                         | 100.0 |
| ICE-HK            |                                                                                   |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        | 162                                                                               | 122                  | 99             | 13        | 57       | 84             | 44            | 63                          | 644   |
| <i>normalised</i> | 32.9                                                                              | 24.8                 | 20.1           | 2.6       | 11.6     | 17.1           | 8.9           | 12.8                        | 131.0 |
| %                 | 25.2                                                                              | 18.9                 | 15.4           | 2.0       | 8.9      | 13.0           | 6.8           | 9.8                         | 100.0 |
| NL                |                                                                                   |                      |                |           |          |                |               |                             |       |
| <i>raw</i>        | 93                                                                                | 54                   | 118            | 9         | 28       | 92             | 58            | 60                          | 512   |
| <i>normalised</i> | 23.2                                                                              | 13.5                 | 29.4           | 2.2       | 7.0      | 22.9           | 14.5          | 15.0                        | 127.6 |
| %                 | 18.2                                                                              | 10.5                 | 23.0           | 1.8       | 5.5      | 18.0           | 11.3          | 11.7                        | 100.0 |

significant). This reminds us of the dangers of using a single ENL variety as *the* native yardstick. It also reminds us that native varieties, too, have been shown to differ from each other in their quantitative preferences (e.g. Mair 2002; Hundt and Vogel 2011), indicating that such differences are not necessarily attributable to varietal type or degree of institutionalisation.

On the whole, the dendrograms of the syntactic and semantic distributions do not validate our hypothesised cline in (3), which reflects norm orientation, nor do they corroborate the opposite cline found for the overall frequency of *into* (Section 3.1), which was attributed to a proficiency effect. This is in line with findings by Laporte (2012), who reports similar heterogeneity in the semantic

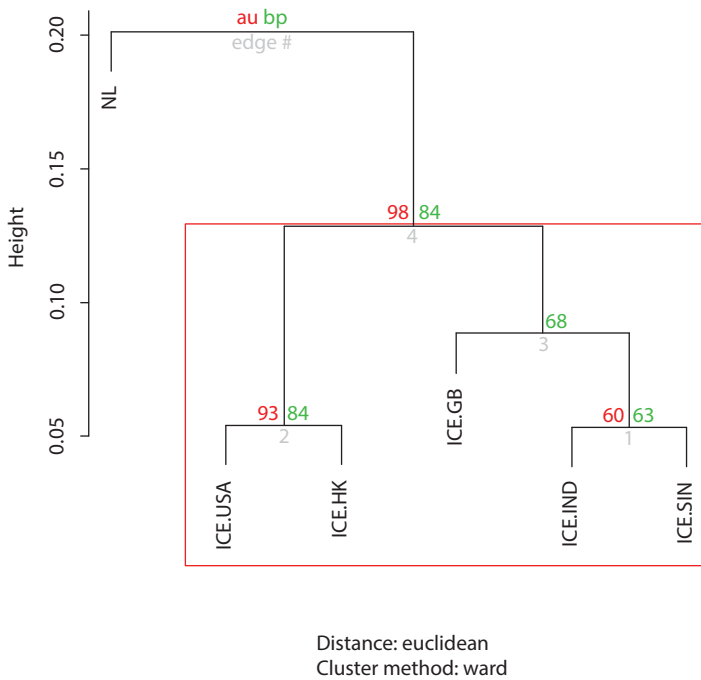


Figure 3. Hierarchical cluster analysis of semantic distribution of *into*

distribution of the verb *make* across ENL, ESL and EFL varieties, and with Gries and Mukherjee's (2010) results from their study of n-grams in ICE-GB, ICE-HK, ICE-IND and ICE-SIN. Gries and Mukherjee (2010) suggest that some phenomena may simply be too fine-grained and too volatile to neatly reflect the degree of institutionalisation of different varieties, a hypothesis reinforced here by the fact that even the two ENL corpora do not pattern together. Varietal type, therefore, does not seem to come into play when it comes to distributional preferences of such specific features.

### 3.2.3 Lexical variation

Table 8 shows the adjusted type/token ratios (TTR)<sup>18</sup> of the verb lemmas used with *into* per corpus. As can be seen — and again contrary to our hypothesised cline — the two most institutionalised varieties, ICE-SIN and ICE-IND, are virtually

18. Gilquin and Granger (2011) report lexical variation in relative frequency of verb types per 100,000 words. As we report TTRs here, we also computed the TTRs for their data for the purposes of comparison. However, it should be noted that TTRs are sensitive to token numbers and thus corpus size: vocabulary growth in corpora being logarithmic, there is bound to be a greater repetition of verbs in a larger corpus, and thus lower lexical variation. As our corpora contain approximately three times as many tokens of *into* as the corpora used in Gilquin and

indistinguishable from ENL. In contrast, the TTRs for the two least institutionalised varieties, ICE-HK and NL, are considerably lower, suggesting that these two varieties tend to rely on a smaller repertoire of verbs with *into* and display a higher degree of repetition of the same verbs.<sup>19</sup> This lower lexical variation, again, seems to be in line with the proficiency effect noted in Gilquin and Granger (2011). The learners in their data had notably low TTRs, ranging from just 34.0 in ICLE-FR to 42.3 in ICLE-TSW, compared to the GB reference corpus of 53.2.<sup>20</sup> It is interesting to note that, as was the case for the overall frequency of *into* (see Section 3.1), there is a considerable gap between the TTR for ICLE-DU in Gilquin and Granger (2011) (39.4) and that for our NL corpus (52.3). As noted previously, the contributors to the latter may be viewed as ‘users’ rather than ‘learners’ of English. Our results thus seem to be determined more by proficiency effects than purported norm orientation.

**Table 8.** Types, tokens and type/token ratio of verb lemmas used with *into* per corpus

| Corpus  | Types | Tokens | TTR  |
|---------|-------|--------|------|
| ICE-GB  | 231   | 583    | 57.6 |
| ICE-USA | 229   | 593    | 57.0 |
| ICE-SIN | 230   | 588    | 57.5 |
| ICE-IND | 209   | 519    | 57.2 |
| ICE-HK  | 194   | 570    | 50.0 |
| NL      | 151   | 440    | 52.3 |

Table 9 shows the 10 most frequent verb lemmas used with *into* in each corpus, their raw frequencies and the cumulative percentages in parentheses. A chi-square test of the sum frequencies of the 10 most frequent verb lemmas in each corpus

---

Granger (2011), we randomly divided all occurrences of *into* in our corpora into three parts and calculated the average TTR to ensure that our results are comparable to theirs.

19. As pointed out by an anonymous reviewer, the lower TTRs with *into* in ICE-HK and NL may reflect an overall lower TTR in these corpora. To verify this, we computed an overall standardised TTR (STTR) using WordSmith Tools 4 for each corpus. The results show that overall STTR indeed varies across corpora, but only marginally so: NL (41.4) < ICE-IND (41.56) < ICE-HK (41.86) < ICE-GB (42.07) < ICE-SIN (42.43) < ICE-USA (43.36). Normalisation of the STTRs against the TTRs of the ‘verbs + *into*’ still yields lower lexical variation for ICE-HK (119.4) and NL (126.3), followed by ICE-USA (131.5), then ICE-SIN (135.5), ICE-GB (136.9) and ICE-IND (137.6). Hence, the overall trend is the same (aside from a slightly lower result for ICE-USA), but for the sake of comparability with Gilquin and Granger’s (2011) results, in the text above we report the TTRs found for the verbs combining with *into*.

20. This is a slightly lower result compared to our ICE-GB and ICE-USA data. As their BrE data consists only of editorials, this may be a text type effect.

returned a highly significant result ( $\chi^2 = 21.81$ , d.f. = 5,  $p < 0.001$ ,  $\phi_c = 0.077$ ), with post-hoc pairwise testing using the Holm correction attributing this to the significant difference between NL and each of the other corpora except ICE-HK.<sup>21</sup> In other words, in the two least institutionalised corpora, NL and ICE-HK, the most frequent verb lemmas account for a greater share of all verbs used with *into* than in the other corpora. This is in line with the TTRs obtained above (see Table 8) and further validates the notion that these corpora tend to rely on a smaller repertoire of verbs. In NL, it is striking that only 4 verbs account for about 25 per cent of all tokens. The top 10 verbs account for over 40 per cent of all verb lemmas used with *into* in NL, approximately 35 per cent in ICE-HK and around 30 per cent in the native corpora as well as ICE-IND and ICE-SIN.

The table also shows that in the native corpora, *get* and *go* are the two most frequent verb lemmas, accounting for close to 10 per cent of all verbs used with *into*. A closer look at the concordances using *get* and *go* in ICE-GB and ICE-USA reveals that they tend to be used in idiomatic/non-compositional ways, such as *go into treatment*, *go into decline* or *get something into line* (e.g. (1) and (2)). The fact that these verbs are considerably less frequent in all NNS varieties may suggest a tendency to make less use of non-compositional expressions with these high-frequency verbs across the ESL and EFL varieties alike.

- (1) It will encourage councils to *get* their budgets **into** line and the voters to throw out those councillors who fail to do that. (ICE-GB W2C-020)
- (2) I'm going to have to set your ms. aside for a while now, so I can get back to work and have our report ready to *go into* print by early '05. (ICE-USA W1B-019)

Also apparent from Table 9 is the fact that, unlike in the native varieties, the verb lemmas *turn* and *take* systematically rank among the top two or three verbs in all NNS varieties, with the exception of *turn* in ICE-IND. *Turn* dominates the transformation category (cf. Table 7) in the NNS varieties (in the sense of *turn into*, e.g. (3) and (4)), whereas the native varieties seem to show higher lexical variation, often encoding the means or manner of the transformation in the verb as well (e.g. (5) and (6)). *Take* appears most frequently in fixed expressions such as *take into account* or *take into consideration* ((7) and (8)). The relatively more frequent use

---

21. The function for this procedure was `chisqPostHoc` (NCStats package for R), which performs pairwise chi square tests for all pairs of corpora, then adjusts the resulting *p*-values for the increased chance of false positives due to multiple pairwise comparisons. The Holm correction was used on the recommendation of an anonymous reviewer. All posthoc chi squares in this paper were calculated in this way.

**Table 9.** Most frequent verb lemmas with *into* per corpus and their cumulative percentages

|    | ICE-GB      |            | ICE-USA     |            | ICE-SIN     |            | ICE-IND     |            | ICE-HK      |            | NL          |            |
|----|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|-------------|------------|
| 1  | <i>get</i>  | 28 (4.8%)  | <i>go</i>   | 30 (5.1%)  | <b>turn</b> | 26 (4.4%)  | <b>take</b> | 24 (4.6%)  | <b>take</b> | 30 (5.3%)  | <b>turn</b> | 41 (9.3%)  |
| 2  | <i>go</i>   | 26 (9.3%)  | <i>get</i>  | 26 (9.4%)  | <b>take</b> | 25 (8.7%)  | look        | 18 (8.1%)  | put         | 30 (10.5%) | <b>take</b> | 31 (16.4%) |
| 3  | come        | 25 (13.6%) | <b>take</b> | 18 (12.5%) | come        | 21 (12.2%) | come        | 16 (11.2%) | <b>turn</b> | 28 (15.4%) | translate   | 23 (21.6%) |
| 4  | <b>turn</b> | 21 (17.2%) | <b>turn</b> | 18 (15.5%) | look        | 19 (15.5%) | <i>go</i>   | 16 (14.3%) | divide      | 26 (20.0%) | <i>go</i>   | 18 (25.7%) |
| 5  | divide      | 19 (20.4%) | put         | 18 (18.4%) | put         | 16 (18.2%) | divide      | 15 (17.1%) | <i>go</i>   | 17 (23.0%) | put         | 15 (29.1%) |
| 6  | <b>take</b> | 19 (23.7%) | transform   | 15 (20.9%) | divide      | 15 (20.7%) | <i>get</i>  | 15 (20.0%) | come        | 17 (26.0%) | come        | 14 (32.3%) |
| 7  | put         | 17 (26.6%) | come        | 15 (23.4%) | transform   | 15 (23.3%) | convert     | 14 (22.7%) | <i>get</i>  | 17 (28.9%) | divide      | 12 (35.0%) |
| 8  | cut         | 11 (28.5%) | fall        | 13 (25.6%) | convert     | 13 (25.5%) | translate   | 12 (25.0%) | fall        | 16 (31.8%) | fit         | 12 (37.7%) |
| 9  | move        | 11 (30.4%) | translate   | 12 (27.6%) | <i>go</i>   | 13 (27.7%) | enter       | 11 (27.2%) | look        | 16 (34.6%) | look        | 11 (40.2%) |
| 10 | transform   | 10 (32.1%) | bring       | 11 (29.6%) | fall        | 13 (29.9%) | <b>turn</b> | 11 (29.3%) | move        | 16 (37.4%) | <i>get</i>  | 9 (42.3%)  |

of *take* among the NNS varieties thus suggests a sort of lexical ‘teddy bear’ effect (Hasselgren 1994) (see also the phraseological uses in Section 3.4 below).

- (3) You can practically *turn* everything that happened **into** stories for him. (ICE-SIN W2D-020)
- (4) The evening had *turned into* Plato’s symposium and my failure was weighing heavily on me: I had forgotten to use the word love! (NL W2F-010)
- (5) It *snapped into* dead-serious mode and didn’t lighten up for days. (ICE-USA W2F-011)
- (6) The basis of every moulding has been defined as the quirk, that is a groove cut into the stonework next to an angle, matched by another parallel quirk, the angle they form then being *smoothed into* a curve. (ICE-GB W2B-003)
- (7) In other words, the financial gain of nuclear energy is not guaranteed if the costs of nuclear waste storage are *taken into account*. (NL W1A-018s1)
- (8) This is an issue on which the views of the people of Hong Kong should be *taken into consideration*. (ICE-HK W2E-009)

In short, while Table 8 showed that ICE-SIN and ICE-IND, the most institutionalised of the NNS varieties under investigation, were indistinguishable from the native varieties in terms of their quantitative variation in the verb lemmas used with *into*, the analysis stemming from Table 9 reveals that lexical parallels can be drawn between the native varieties on the one hand and the NNS varieties on the other.

### 3.2.4 *Phraseological uses*

In this section we zoom in on the (semi-)fixed expressions (SFEs) category from the semantic classification in Section 3.2.2. Table 7 showed that the proportion of SFEs using *into* ranges from a low of 6.9 per cent in ICE-USA to highs of 9.8 per cent in ICE-HK and 11.7 per cent in NL, though the differences were not significant. This trend towards higher use of SFEs in the least institutionalised varieties is in line with Gilquin and Granger’s (2011:66) findings, where the learners with the least exposure (French- and Spanish-speaking learners) showed the highest frequency of SFEs. According to Gilquin and Granger (2011:66), this “seems to contradict the common claim that a great deal of exposure is necessary in order to acquire formulaic expressions”. They posit two factors that may underlie this: (a) the tendency of learners to stick to expressions they know and consider safe, and (b) positive transfer. Both of these factors are corroborated by our data.

First, Table 10 below shows the SFEs that occurred at least five times in any corpus and their cumulative percentages. ICE-HK and NL, the least institutionalised

varieties, have more SFEs that are used five times or more than the other corpora. Moreover, a chi-square test of the cumulative frequencies for the SFEs that appear five or more times per corpus returned a highly significant result ( $\chi^2 = 22.24$ ,  $df = 5$ ,  $p < 0.001$ ,  $\phi_c = 0.267$ ), which post-hoc testing attributed to the significant pairwise differences between NL and ICE-USA and between ICE-HK and ICE-USA. In other words, there is a significantly higher degree of reliance on a select repertoire of common SFEs in ICE-HK and NL. It is worth noting, however, that our proportion of SFEs for NL (11.7 per cent, cf. Table 7), though higher than that for all the other present corpora, is still much lower than Gilquin and Granger's (2011) result for ICLE-DU (25.5 per cent), lending further support to the notion of a proficiency effect — the more advanced the English level, the less reliance on fixed chunks.

**Table 10.** Most frequent (semi-)fixed expressions per corpus and their cumulative percentages

| Corpus  | Expression              | Freq. and cum. % |
|---------|-------------------------|------------------|
| ICE-GB  | take into account       | 15 (27.8%)       |
|         | come into play          | 6 (38.9%)        |
| ICE-USA | take into account       | 8 (17.4%)        |
| ICE-SIN | take into account       | 9 (19.1%)        |
|         | take into consideration | 9 (38.2%)        |
| ICE-IND | take into consideration | 10 (23.3%)       |
|         | take into account       | 9 (44.2%)        |
| ICE-HK  | take into account       | 15 (23.8%)       |
|         | take into consideration | 10 (39.7%)       |
|         | be into                 | 5 (47.6%)        |
| NL      | take into account       | 27 (45.0%)       |
|         | be into                 | 5 (53.3%)        |
|         | come into contact       | 5 (61.7%)        |

Second — with regard to positive transfer — the SFE *take into account* and its synonymous counterpart *take into consideration* warrant further discussion. Although the figures are relatively low, Table 11 seems to show a clear preference in the native varieties and NL for *take into account*, while ICE-HK, and ICE-IND and ICE-SIN display more even distributions of the two expressions. Nesselhauf (2009) reports the same preference for *take into account* in BrE in the BNC, and similarly even proportions between the two expressions in four Outer Circle varieties (Jamaican, Kenyan, Indian and Singapore English). However, she also reports even distributions for four Expanding Circle varieties (with the L1s Polish, German, French and Finnish), while here NL, with a markedly higher use of *take into account*, shows the opposite trend. Nesselhauf (2009) attributes the higher use of *take into*

*consideration* in NNS varieties to the fact that it is more regular intralinguistically due to the related verb *consider*, which has roughly the same meaning, and because it is less idiomatic than *take into account*. Language internal (ir)regularities could be at the root of similar processes in non-native varieties that give rise to parallels across both EFL and ESL varieties. However, the results for NL seem to contradict this. The most likely explanation for this opposite trend in NL may be that of positive transfer: the equivalent Dutch expression *rekening houden met* is lexically closer to *take into account* as the noun *rekening* translates directly into *account*. This illustrates the intricate interplay between shared tendencies stemming from language internal (ir)regularities and L1 influence that accounts for pockets of idiosyncrasy in some varieties. While there is clearly a case for some common processes occurring across NNS varieties, it seems that for NL the salience of the L1 equivalent outweighs the effect of language internal constellations.

Table 11. Raw frequencies of *take into account* versus *take into consideration* per corpus

|                                | ICE-GB | ICE-USA | ICE-SIN | ICE-IND | ICE-HK | NL |
|--------------------------------|--------|---------|---------|---------|--------|----|
| <i>take into account</i>       | 15     | 8       | 9       | 9       | 15     | 27 |
| <i>take into consideration</i> | 1      | 2       | 9       | 10      | 10     | 3  |

### 3.2.5 Nonstandard uses

Table 12 shows the results of the classification of nonstandard uses of *into*. The ENL corpora have the lowest percentages of nonstandard uses, as may be expected (2.6 per cent each). ICE-SIN and ICE-IND have 5.0 per cent and 6.0 per cent respectively, closely followed by NL (6.4 per cent), with ICE-HK trailing at 9.2 per cent. A chi square test returned a highly significant result ( $\chi^2 = 42.79$ ,  $df = 5$ ,  $p < 0.001$ ,  $\phi_c = 0.107$ ), with post-hoc pairwise testing attributing this to the significant proportions of nonstandard uses in ICE-HK, NL and ICE-IND compared to the two ENL corpora. ICE-HK also has significantly more nonstandard uses than ICE-SIN. No significant differences were found between ICE-SIN and the ENL corpora. In other words, here again, the ESL varieties follow the opposite cline than that hypothesised in (3): the more advanced the variety in Schneider's model, the more similar it is to the ENL varieties. In comparison, Gilquin and Granger (2011) found a low of 4.7 per cent of nonstandard uses for ICLE-FR, which they suggested could be due to a play-it-safe strategy, followed by 6.5 per cent for ICLE-DU, which is just slightly higher than our result for NL (6.4 per cent). They also found considerably higher percentages of nonstandard uses for ICLE-SP and ICLE-TSW (15.8 per cent and 30.5 per cent, respectively), which they attributed to less attention to form during instruction in ICLE-TSW and less exposure to English as well as negative transfer in ICLE-SP.

Table 12. Nonstandard uses of *into* per corpus

| Corpus  | Raw frequency | %   |
|---------|---------------|-----|
| ICE-GB  | 17            | 2.6 |
| ICE-USA | 17            | 2.6 |
| ICE-SIN | 32            | 5.0 |
| ICE-IND | 34            | 6.0 |
| ICE-HK  | 59            | 9.2 |
| NL      | 33            | 6.4 |

Taking a more qualitative perspective, the nonstandard uses of *into* seem to fall into four (sometimes overlapping) categories.

1. The use of *into* where the simpler form *in* or *to* would conventionally be preferred. Many of these forms, such as *elect into*, *include into*, *admit into* and *add into*, were found across three or more varieties, suggesting wide use.
  - (9) In chapter II, I will therefore propose a new way of proceeding by *including* the role of the small power **into** our research agenda. (NL W1A-006)
  - (10) At an exhibition put up by 48 Australian educational institutions here, they ask the universities to name their brightest academics as one way of checking whether they would be *enrolling into* a good university. (ICE-SIN W2C-002)
  - (11) Ever increasing urbanization, surging population, ramifying industries, the green revolution and their aftermath have *culminated into* an altogether different type of problem termed as the environmental pollution. (ICE-IND W2A-037)
2. Hyper-explicitness, often manifesting itself in the use of *into* where conventionally a null form would be preferred, e.g. *enter into* rather than simply *enter* in the concrete sense of ‘a room’. This is in line with Nesselhauf (2009), who observed that the direction already expressed in verbs of movement tends to be made more explicit in ESL and EFL varieties; for instance, in cases such as *enter into*, *approach to* and *return back*, the directionality is already expressed in the verb. The case of *enter into* seems to arise from analogy with *go into* and perhaps also with standard forms such as *enter into a partnership/discussions*. The examples in (13) and (14) could also fall under the first category above as they use *into* instead of the simpler forms *in* and *to*, but in these cases the emphasis seems to be on making explicit the direction/movement involved.
  - (12) The larva *enter into* the soil and it pupates into the soil below 10 cm. (ICE-IND W1A-019)

- (13) All the money that is allocated to these events will *land into* the coffers of big business, FIFA and the government. (NL W2C-019s1)
- (14) When starch granules are *added into* cold water, water molecules can freely penetrate into the starch granule. (ICE-HK W1A-018)
3. Variation on existing patterns, where the preposition *into* is unconventionally used in an otherwise standard expression. For example, the standard prepositions in (15)–(17) and (18) below would typically be *to* and *in*, respectively. These examples could also fall under the first category as they use *into* instead of simpler forms.
  - (15) Bhajan Sopori is one of the leading santoor players who have *brought this* ancient, 100-string instrument *into prominence* during the last 20 years. (ICE-IND W2D-013)
  - (16) Another cautioned me, “Don’t *jump into conclusions* so hastily”. (ICE-SIN W2F-005)
  - (17) It is a process in which each person learns to *put* his biological, physical tools *into use* according to the rules and norms of one’s culture and society. (NL W2A-014)
  - (18) Now it seems as if mentally retarded are natural zen-masters. In comparison to us, they appear to be spontaneous, reacting freely to what is instead of what they judge the situation to be. They are *into the moment*, concentrated, in flow all the time, as it were. (NL W2B-005)
4. Blends of existing patterns. This phenomenon was also reported by Gilquin and Granger (2011), with examples such as *take into seriousness* as a blend of *take seriously* and *take into consideration*. In the present data, numerous examples of (19) were found in ICE-HK, where *take into account of* seems to arise from a combination of *take into account* and *take account of*. Similarly, in (20) *land into trouble* could be a blend of *get into trouble* and *land in hot water*. In this case, the use of *into* instead of *in* again seems to emphasise movement or directionality, as in category 2 above.
  - (19) In the proposed new process, it states that the planning Board (PB) should *take into account of* the results and findings of planning studies, public opinions received, and relevant government policies in drawing up detailed planning proposals. (ICE-HK W2A-020)
  - (20) Atul’s tendency of worrying too much invariably *landed him into trouble* time and again. (ICE-IND W2F-013)

What should be noted here is that, as in Nesselhauf (2009), we find the same or similar nonstandard structures being used not just across varieties, but also across varietal types, particularly across both Outer and Expanding Circle varieties. In addition, we find similar types of nonstandard uses as those in Gilquin and Granger's (2011) learner data. This can be illustrated with the analogous examples of *result into* as in (21), from their data, and (22), from our ESL data. Such findings provide support for the notion of a continuum rather than a strict divide between ESL and EFL varieties.

- (21) This *resulted into* one scarverging for employment in other the cope with advanced life in the city. At the end it encourages prostitution. (ICLE-TSW, Gilquin and Granger 2011)
- (22) Technological advancement, particularly, the rapid computerisation, depersonalised the work, restricting the scope for human interaction. This *resulted into* a deep sense of growing loneliness which affected the individual life. (ICE-IND W2A-005)

#### 4. Discussion

This paper aimed to provide an integrated analysis of the preposition *into* in varieties spanning all three of Kachru's (1985) circles. We set out to investigate previous suggestions that the more advanced a variety is in Schneider's (2003, 2007) Dynamic Model, the more dissimilar it would be from ENL. To provide an overview of the findings, Table 13 compares the hypothesised cline with the actually observed clines for each of the variables under investigation. It shows that the results for ICE-GB and ICE-USA are heterogeneous; rather than always appearing together in the cline they are frequently interspersed with ICE-SIN or ICE-IND. These latter two varieties — the most institutionalised of the ESL varieties under investigation — perform consistently more similarly to the ENL varieties than the least institutionalised varieties, NL and ICE-HK. The trend therefore runs counter to the hypothesis: rather than the most norm-dependent varieties performing in the most exonormative and the most norm-developing varieties in the most dissimilar way to the ENL varieties, we found the reverse. The more advanced a variety in Schneider's model (i.e. the more institutionalised), the more similar it was to ENL, while the least institutionalised varieties were the most distant from ENL. It would seem that, at least for the use of the preposition *into*, a purportedly endonormative status (ICE-SIN and ICE-IND) does not necessarily equate to dissimilarity from ENL varieties, and a more exonormative orientation (ICE-HK and NL) does not necessarily guarantee more native-like performance. Thus a distinction

should be drawn between attitudinal orientation and actual performance, whereby factors such as proficiency and exposure can outweigh norm orientation.

**Table 13.** Overview of observed cline per variable

| Variable                                                                     | Hypothesised cline                 |
|------------------------------------------------------------------------------|------------------------------------|
|                                                                              | USA/GB — NL — HK — IND — SIN       |
|                                                                              | Observed cline                     |
| Overall frequencies                                                          | {SIN — USA — GB} — {IND — HK — NL} |
| Syntactic distribution                                                       | {SIN} — {USA/GB/HK} — {NL/IND}     |
| Semantic distribution                                                        | {USA/HK/GB/IND/SIN} — NL           |
| Lexical variation                                                            |                                    |
| TTR                                                                          | GB — SIN — IND — USA — NL — HK     |
| Cumulative % of most frequent verb lemmas                                    | {IND — USA — SIN — GB} — {HK — NL} |
| Phraseological uses (cumulative % of most frequent (semi-)fixed expressions) | USA — SIN — GB — IND — HK — NL     |
| Nonstandard uses                                                             | {USA — GB — SIN} — IND — NL — HK   |

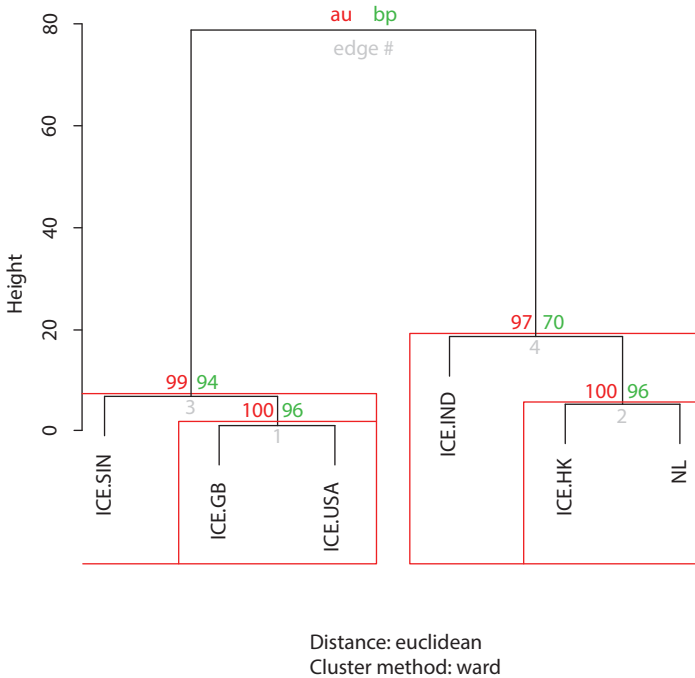
*Note:* Curly brackets denote clusters of varieties that did not differ significantly from one another.

The findings suggest that, as the least institutionalised varieties, ICE-HK and NL display more learner-like characteristics than ICE-SIN and ICE-IND (though not to the extent of the ‘true’ learners in Gilquin and Granger (2011), as will be discussed below). Both ICE-HK and NL had a significantly lower frequency of *into* overall (Table 5, Figure 1). With regard to lexical variation in the verbs used with *into*, ICE-HK and NL had lower TTRs (cf. Table 8), suggesting that these two varieties rely on a smaller repertoire of verbs with a higher degree of repetition. Further, the most frequent verb lemmas used with *into* in ICE-HK and NL accounted for a greater share of all verb lemmas used with *into* than in the other corpora (Table 9). ICE-HK and NL also had the highest frequencies of (semi-) fixed expressions, though the differences were not significant. Lastly, in the non-standard classification (Table 12), NL and in particular ICE-HK had the highest proportions of nonstandard uses of *into*.

These findings are in line with the notion of a continuum rather than a strict divide between varietal types, where some Outer Circle varieties, like HKE, may be less prototypically ESL, and some Expanding Circle varieties, like NL, may be less prototypically EFL (see e.g. Biewer 2011: 27–28). Along with our more heterogeneous findings for the semantic and syntactic classification, whereby the distributional preferences in Figure 2 and Figure 3 seem to have little relation with varietal type or degree of institutionalisation, this provides support for recent claims by Davydova (2012: 384), Hundt and Vogel (2011) and Werner (2013) at

Schneider’s model — and categorisations such as ESL versus EFL — may be more applicable when considering sociocultural aspects such as identity issues, but less so for investigating structural features in isolation. It also ties in with Gries and Mukherjee’s (2010) assertion for the phenomenon of n-grams, that some linguistic features may be simply too fine-grained and variable to reflect sociocultural aspects such as norm orientation.

This is corroborated by Figure 4, which shows the output of a hierarchical cluster analysis based on all the variables we investigated.<sup>22</sup> ICE-SIN clusters with the native corpora, leading us to reconsider the strict divide often claimed to exist between ESL and ENL varieties. On the adjacent branch, NL and ICE-HK cluster closely, together with, to a lesser extent, ICE-IND. In other words, NL does not perform markedly differently from the ESL corpora, but instead shares similar patterns of divergence with respect to the ENL varieties. Where the quantitative



**Figure 4.** Hierarchical cluster analysis of all variables

22. Excluding the syntactic and semantic distributions, as these were considered separately in Figures 2 and 3. The input data for each variable is expressed in frequency per 100,000 words (see Appendix 3 for the input data). This measure is slightly different from the TTR reported for lexical variation in Section 3.3 and the percentage reported for nonstandard uses in Section 3.5. However, this has no impact on the conclusions regarding the lexical variation and nonstandard uses reported above and allows the same scale to be used for all three variables included in the HCA.

findings largely show a split between ICE-SIN and ICE-IND on the one hand and ICE-HK and NL on the other, qualitative analyses tend to show similar behaviour across all four NNS varieties, in contrast to the two native varieties. For example, in terms of lexical variation, ICE-GB and ICE-USA show systematic preferences for *get* and *go*, which are typically used in non-compositional ways such as *to go into treatment*, whereas the NNS varieties seem to prefer fixed expressions like *take into account* or *take into consideration*. In terms of nonstandard uses all NNS varieties show similar types of divergence from Standard English, such as hyper-explicitness (e.g. *enter into* in the concrete sense of entering a room) and variation on existing patterns (e.g. *jump into conclusions*, *put into use*). Indeed, some of the nonstandard uses identified in our ESL data are exactly the same as those in Gilquin and Granger's (2011) learner data, highlighting the fact that these processes cut across both ESL and EFL varietal types. To this end they pointed out that "[t]he line is thin between errors and creative uses", and that maintaining a strict divide between purportedly norm-developing ESL varieties and norm-dependent EFL varieties, paradoxically, "amounts to interpreting departures from native standards as errors in the case of learner English and as creative innovations in the case of institutionalised L2 varieties" (Gilquin and Granger 2011: 72). This also ties in with Nesselhauf (2009), who found the same types of new prepositional verbs in both ESL (ICE) and EFL (ICLE) data, as well as other authors who discussed similar extended uses across varietal types (Edwards 2011; Erling 2002: 10; Laporte 2012: 284; Schneider 2012: 70). What is more, the native varieties in the present data show comparable nonstandard uses as the NNS varieties, such as *credit into* and *admit into*. That these processes seem to be present across all three varietal types provides further support for the notion of a continuum as discussed above, with the corollary that varietal types should be approached in an integrated fashion (Buschfeld 2011; Davydova 2012; Mukherjee and Hundt 2011).

The development of the NL corpus (Edwards 2011, [E]) with the full range of ICE written text types, has enabled us to explore all three varietal types using data that are, crucially, not restricted just to learner writing. This also allows us to compare our results for 'Dutch English' from the NL corpus with Gilquin and Granger's (2011) findings from ICLE-DU. The NL corpus includes texts by academics, journalists and other professionals, whereas ICLE-DU includes learners only. The results reveal the importance of differentiating between 'users' of English and 'learners' proper. For example, the lexical variation in the verb lemmas used with *into* was considerably higher in NL than in ICLE-DU (52.3 per cent and 39.4 per cent, respectively), pointing to overall higher proficiency among the NL corpus contributors compared to the learners in ICLE-DU. Similarly, NL showed far less reliance on SFEs than ICLE-DU (11.7 per cent versus 25.5 per cent, respectively), again suggesting a proficiency effect. This distinction between NL

and ICLE-DU is highlighted in Figure 5, which is based on the same quantitative variables as Figure 4 above (see Appendix 4 for the input data). The cluster on the right is the same as in Figure 4, using data from the present study, while the cluster on the left is derived from the ICLE data in Gilquin and Granger (2011). The figure shows a clear split between the ‘learners’ on the left and the ‘users’ on the right, most obviously in that ICLE-DU clusters with the other ICLE corpora, while NL clusters with the ICE varieties.<sup>23</sup> This clearly demonstrates the need for caution — and, crucially, comparable corpus data — when making claims about systematic differences between ESL and purported EFL varieties.

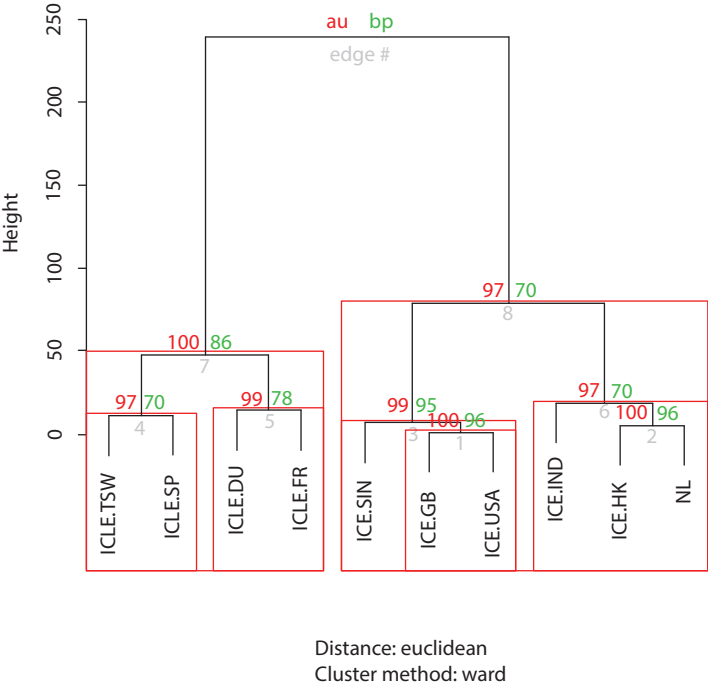


Figure 5. Hierarchical cluster analysis of all variables (ICE versus ICLE)

23. It should be noted that the variable COUNTRY OF ORIGIN may also play a role when comparing the ICLE-DU and the NL corpus; while the former comprises texts from both Dutch and Belgian contributors, the latter includes only Dutch contributors. In the Netherlands, pupils start learning English compulsorily at age 10. In Belgium, pupils can choose between taking French, German or English as their first foreign language but in Dutch-speaking Belgium, a majority of 60 per cent opts for English (typically starting around age 10, or if only opted for in high school, starting between the ages 12 and 14) (Meunier 2009: 107). Both countries subtitle rather than dub television, thus all in all, children have a similar amount of exposure, but sometimes start formal learning of English somewhat later in Dutch-speaking Belgium than in the Netherlands.

## 5. Conclusion

We investigated the patterning of the preposition *into* in English varieties spanning all three of Kachru's (1985) circles, including the first Expanding Circle corpus that includes the full range of ICE written text types (the Corpus of Dutch English, NL). The quantitative results showed that the more institutionalised ESL varieties were the most similar to ENL, suggesting that 'norm-developing' ESL status need not preclude an ongoing exonormative orientation; while the least institutionalised varieties were the most dissimilar to ENL, indicating that an exonormative orientation is not the only factor at play. The qualitative analyses showed similarities that cut across the EFL and ESL boundary, as well as the EFL, ESL and ENL boundaries, indicating similar processes at work across Kachruvian circles. Moreover, the NL data did not differ markedly from the ESL data, leading us to question the strict divide held to exist between varietal types. However, the findings for NL were notably different to those for ICLE-DU (Gilquin and Granger 2011), suggesting that the Expanding Circle may be home to 'users' as well as 'learners'. This highlights the importance of developing comparable, wide-ranging corpora wherever possible and appropriate, to do justice to the expanding uses of English in today's Expanding Circle.

## References

- Ammon, Ulrich, and Grant McConnell. 2002. *English as an Academic Language in Europe*. Frankfurt: Peter Lang.
- Biewer, Carolin. 2011. "Modal Auxiliaries in Second Language Varieties of English: A Learner's Perspective". In Joybrato Mukherjee and Marianne Hundt, eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 7–33. DOI: 10.1075/scl.44.02bie
- Booij, Geert. 2001. "English as the Lingua Franca of Europe: A Dutch Perspective". *Lingua e Stile* 36: 347–357.
- Buschfeld, Sarah. 2011. *The English Language in Cyprus: An Empirical Investigation of Variety Status*. PhD dissertation, University of Cologne.
- Davydova, Julia. 2012. "Englishes in the Outer and Expanding Circles: A Comparative Study". *World Englishes*, 31: 366–385. DOI: 10.1111/j.1467-971X.2012.01763.x
- De Cock, Sylvie, and Sylviane Granger. 2004. "High Frequency Words: The Bête Noire of Lexicographers and Learners Alike. A Close Look at the Verb Make in Five Monolingual Learners' Dictionaries of English". In Geoffrey Williams and Sandra Vessier, eds. *Proceedings of the 11th EURALEX International Congress*. Lorient: University of Southern Brittany, 233–243.
- Edwards, Alison. 2011. *English in the Netherlands: Functions, Forms and Attitudes*. PhD dissertation, University of Cambridge.
- Edwards, Alison. 2011. "Introducing the Corpus of Dutch English". *English Today* 27: 10–14. DOI: 10.1017/S0266078411000319

- Edwards, Alison. 2011. "The EFL–ESL Continuum and the Case of the Netherlands: A Comparative Analysis of the Progressive Aspect". *World Englishes* 33: 173–94. DOI: 10.1111/weng.12080
- Erling, Elizabeth J. 2002. "'I Learn English Since Ten Years': The Global English Debate and the German University Classroom". *English Today* 18: 8–13. DOI: 10.1017/S026607840200202X
- European Commission. 2012. *Europeans and Their Languages: Special Eurobarometer* 386. Brussels.
- Gilquin, Gaëtanelle, and Sylviane Granger. 2011. "From EFL to ESL: Evidence from the International Corpus of Learner English". In Joybrato Mukherjee and Marianne Hundt, eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 55–78. DOI: 10.1075/scl.44.04gra
- Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *The International Corpus of Learner English. Version 2*. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.
- Greenbaum, Sidney. 1991. "ICE: The International Corpus of English". *English Today* 7: 3–7. DOI: 10.1017/S0266078400005836
- Gries, Stefan Th. 2010. "Behavioral Profiles: A Fine-Grained and Quantitative Approach in Corpus-Based Lexical Semantics". *The Mental Lexicon* 5: 323–346. DOI: 10.1075/ml.5.3.04gri
- Gries, Stefan Th., and Joybrato Mukherjee. 2010. "Lexical Gravity Across Varieties of English: An ICE-Based Study of N-Grams in Asian Englishes". *International Journal of Corpus Linguistics* 15: 520–548. DOI: 10.1075/ijcl.15.4.04gri
- Hasselgren, Angela. 1994. "Lexical Teddy Bears and Advanced Learners: A Study into the Ways Norwegian Students Cope with English Vocabulary". *International Journal of Applied Linguistics* 4: 237–258. DOI: 10.1111/j.1473-4192.1994.tb00065.x
- Hofland, Knut, and Stig Johansson. 1982. *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.
- Holmes, Janet. 1996. "The New Zealand Spoken Component of ICE: Some Methodological Challenges". In Sidney Greenbaum, ed. *Comparing English World-Wide: The International Corpus of English*. Oxford: Oxford University Press, 163–181.
- Hundt, Marianne, and Katrin Vogel. 2011. "Overuse of the Progressive in ESL and Learner Englishes – Fact or Fiction?" In Joybrato Mukherjee and Marianne Hundt, eds. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 145–166. DOI: 10.1075/scl.44.08vog
- Jenkins, Jennifer. 2009. *World Englishes: A Resource Book for Students* (2nd ed.). Oxon: Routledge.
- Kachru, Braj. B. 1985. "Standards, Codification and Sociolinguistic Realism: The English Language in the Outer Circle". In R. Quirk and H. Widdowson, eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.
- Kirkpatrick, Andy. 2007. *World Englishes: Implications for International Communication and English Language Teaching*. Cambridge: Cambridge University Press.
- Laitinen, Mikko. 2011. "Contacts and Variability in International Englishes: Compiling and Using the Corpus of English in Finland". *Studies in Variation, Contacts and Change in English* 6. Retrieved from [www.helsinki.fi/varieng/series/volumes/06/](http://www.helsinki.fi/varieng/series/volumes/06/)
- Laporte, Samantha. 2012. "Mind the Gap! Bridge between World and Learner Englishes in the Making". *English Text Construction* 5: 264–291. DOI: 10.1075/etc.5.2.05lap

- Mahboob, Ahmar. 2008. "Pakistani English: An Overview of Its Syntax, Morphology, and Lexis". In Rajend Mesthrie, ed. *A Handbook of Varieties of English*, Vol. 4. Berlin and New York: Mouton de Gruyter.
- Mair, Christian. 2002. "Three Changing Patterns of Verb Complementation in Late Modern English". *English Language and Linguistics* 6: 105–132. DOI: 10.1017/S1360674302001065
- McArthur, Tom. 1998. *The English Languages*. Cambridge: Cambridge University Press.
- Mesthrie, Rajend, and Rakesh Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.
- Meunier, Fanny. 2009. "The Status of English in Belgium". In Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot, eds. *The International Corpus of Learner English. Version 2. Handbook and CD-Rom*. Louvain-la-Neuve: Presses Universitaires de Louvain, 106–113.
- Mukherjee, Joybrato, and Marianne Hundt. 2011. "Introduction". In *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 1–5. DOI: 10.1075/scl.44
- Mukherjee, Joybrato, and Stefan Th. Gries. 2009. "Collostructional Nativisation in New Englishes: Verb-Construction Associations in the International Corpus of English". *English World-Wide* 30: 27–51. DOI: 10.1075/eww.30.1.03muk
- Mukherjee, Joybrato. 2009. "The Lexicogrammar of Present-Day Indian English: Corpus-Based Perspectives on Structural Nativisation". In Ute Römer and Rainer Schulze, eds. *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins, 117–135. DOI: 10.1075/scl.35.9muk
- Mukherjee, Joybrato, and Marianne Hundt. 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins. DOI: 10.1075/scl.44
- Mwangi, Serah. 2003. *Prepositions in Kenyan English: A Corpus-Based Study in Lexico-Grammatical Variation*. Aachen: Shaker Verlag.
- Nesselhauf, Nadja. 2009. "Co-Selection Phenomena Across New Englishes: Parallels (and Differences) to Foreign Learner Varieties". *English World-Wide* 30: 1–26. DOI: 10.1075/eww.30.1.02nes
- Peters, Pam. 1998. "In Quest of International English: Mapping the Levels of Regional Divergence". In Antoinette Renouf, ed. *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 281–292.
- Quirk, Randolph, Jan Svartvik, Geoffrey Leech, and Sidney Greenbaum. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.
- Schneider, Edgar W. 2003. "The Dynamics of New Englishes: From Identity Construction to Dialect Birth". *Language*, 79: 233–281. DOI: 10.1353/lan.2003.0136
- Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511618901
- Schneider, Edgar W. 2012. "Exploring the Interface between World Englishes and Second Language Acquisition – and Implications for English as a Lingua Franca". *Journal of English as a Lingua Franca* 1: 57–91. DOI: 10.1515/jelf-2012-0004
- Suzuki, Ryota, and Hidetoshi Shimodaira. 2006. "pvclust: An R Package for Hierarchical Clustering with p-values". *Bioinformatics* 22: 1540–1542. DOI: 10.1093/bioinformatics/btl117
- Van Rooy, Bertus. 2006. "The Extension of the Progressive Aspect in Black South African English". *World Englishes* 25: 37–64. DOI: 10.1111/j.0083-2919.2006.00446.x

Werner, Valentin. 2011. Temporal Adverbials and the Present Perfect/Past Tense Alternation”. *English World-Wide* 34.2: 202–240. DOI: 10.1075/eww.34.2.04wer

Werner, Valentin. 2011. *The Present Perfect in World Englishes: Charting Unity and Diversity*. PhD dissertation, University of Bamberg.

Appendix 1: Input data for Figure 2: Hierarchical cluster analysis of syntactic distribution of *into*

| Syntactic structure                | ICE-GB | ICE-USA | NL    | ICE-HK | ICE-IND | ICE-SIN |
|------------------------------------|--------|---------|-------|--------|---------|---------|
| V <sub>trans</sub> + <i>into</i>   | 0.499  | 0.495   | 0.449 | 0.508  | 0.465   | 0.535   |
| V <sub>intrans</sub> + <i>into</i> | 0.377  | 0.395   | 0.410 | 0.377  | 0.453   | 0.390   |
| Noun + <i>into</i>                 | 0.108  | 0.098   | 0.125 | 0.107  | 0.074   | 0.072   |
| Other                              | 0.015  | 0.011   | 0.016 | 0.008  | 0.007   | 0.003   |

Appendix 2: Input data for Figure 3: Hierarchical cluster analysis of semantic distribution of *into*

| Semantic class           | ICE-GB | ICE-USA | NL    | ICE-HK | ICE-IND | ICE-SIN |
|--------------------------|--------|---------|-------|--------|---------|---------|
| movement                 | 0.305  | 0.251   | 0.182 | 0.252  | 0.278   | 0.281   |
| abstract movement        | 0.116  | 0.200   | 0.105 | 0.189  | 0.154   | 0.142   |
| transformation           | 0.141  | 0.165   | 0.230 | 0.154  | 0.198   | 0.187   |
| causation                | 0.047  | 0.030   | 0.018 | 0.020  | 0.030   | 0.022   |
| division                 | 0.081  | 0.054   | 0.055 | 0.089  | 0.085   | 0.072   |
| other meanings           | 0.137  | 0.147   | 0.180 | 0.130  | 0.081   | 0.129   |
| phrasal verbs            | 0.092  | 0.083   | 0.113 | 0.068  | 0.097   | 0.093   |
| (semi-)fixed expressions | 0.081  | 0.069   | 0.117 | 0.098  | 0.076   | 0.074   |

Appendix 3: Input data for Figure 6: Hierarchical cluster analysis of all variables (frequencies per 100,000 words)

| Variable          | ICE-GB | ICE-USA | NL    | ICE-HK | ICE-IND | ICE-SIN |
|-------------------|--------|---------|-------|--------|---------|---------|
| Frequency         | 157.4  | 158.6   | 127.6 | 131.0  | 137.3   | 159.3   |
| Lexical variation | 68.4   | 68.7    | 48.4  | 50.1   | 62      | 72.2    |
| Non-standard uses | 4.0    | 4.1     | 8.2   | 12.0   | 8.3     | 8.0     |

**Appendix 4: Input data for Figure 7: Hierarchical cluster analysis of all variables (ICE vs ICLE) (frequencies per 100,000 words)**

| Variable          | ICE-GB | ICE-USA | ICE-SIN | ICE-IND | ICE-HK | NL   | ICLE-DU | ICLE-FR | ICLE-TSW | ICLE-SP |
|-------------------|--------|---------|---------|---------|--------|------|---------|---------|----------|---------|
| Frequency         | 157.4  | 158.6   | 159.3   | 137.3   | 131.0  | 12.6 | 103.8   | 92.7    | 70.7     | 72.7    |
| Lexical variation | 68.4   | 68.7    | 72.2    | 62      | 50.1   | 48.4 | 38.4    | 29.6    | 29.1     | 24.9    |
| Non-standard uses | 4.0    | 4.1     | 8.0     | 8.3     | 12.0   | 8.2  | 6.7     | 4.4     | 21.6     | 11.5    |

*Authors' addresses*

Alison Edwards  
 Department of Theoretical and Applied  
 Linguistics  
 University of Cambridge  
 584 King's College  
 Cambridge CB2 1ST  
 UK  
 ae302@cam.ac.uk

Samantha Laporte  
 Faculté de philosophie, arts et lettres  
 Université catholique de Louvain  
 Place Blaise Pascal 1, L3.03.33  
 1348 Louvain-la-Neuve  
 Belgium  
 samantha.laporte@uclouvain.be