Lexical gravity across varieties of English

An ICE-based study of *n*-grams in Asian Englishes*

Stefan Th. Gries and Joybrato Mukherjee
University of California, Santa Barbara / Justus Liebig University, Giessen

In our earlier work on three Asian Englishes and British English, we showed how lexico-syntactic co-occurrence preferences for three argument structure constructions revealed differences between varieties that correlated well with Schneider's (2003, 2007) model of evolutionary stages. Here, we turn to lexical co-occurrence preferences and investigate if and to what degree *n*-grams distinguish between different modes and varieties in the same components of the International Corpus of English. Our approach to *n*-grams differs from previous work in that we neither use raw frequencies nor (problematic) *MI*-values but the newly proposed measure of lexical gravity (cf. Daudaravičius & Marcinkevičienė 2004), which takes type frequencies into consideration. We show how lexical gravity can be extended to handle *n*-grams with $n \geq 3$ and apply this method to our *n*-gram data; in addition, we suggest a new concept for describing the tendency of a word to occur in significant *n*-grams: lexical stickiness.

Keywords: lexical gravity, *n*-gram, lexical stickiness, Asian Englishes, modes

1. Introduction

1.1 *N*-grams in today's corpus linguistics

One particularly attractive method in corpus linguistics these days is the study of *n*-grams, i.e. (usually) uninterrupted sequences of *n* words, and it is in this sense that we will use the term '*n*-gram' in the present paper. In some sense at least, the current prominence of this method comes as no surprise: First, recent corpus developments, technological advances, and the more widespread use of programming languages in corpus linguistics have made the retrieval and analysis of *n*-grams much easier and much more promising (although quite a few methodological problems still await resolution; cf. below). Second, it is precisely in the area of *n*-grams where corpus linguistics begins to enter into a closer relationship with

International Journal of Corpus Linguistics 15:4 (2010), 520–548. DOI 10.1075/ijcl.15.4.04gri
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

the neighboring disciplines of cognitive linguistics and psycholinguistics (as, for example, discussed by Gries 2010b, and Mukherjee 2010). For example,

- Bell et al. (2003) discuss how words are shorter to produce when they are part of a more frequent 2-gram or 3-gram;
- Underwood et al. (2004) show that subjects need fewer eye fixations to read formulaic sequences that are up to six words long;
- McDonald & Shillcock (2003) and Reali & Christiansen (2007) demonstrate that words are faster to process when they appear as part of a more frequent 2-gram;
- Bannard & Matthews (2008) show that children as young as two and three years old are faster and more accurate at repeating high-frequency phrases compared to lower-frequency phrases even when part frequency is controlled for;
- Arnon & Snider (2010) demonstrate that subjects process more frequent 4-grams faster than less frequent ones, etc.

Thus, there is a lot of evidence that the very basic corpus-linguistic concept of *n*-grams has some kind of psychological reality in the sense that *n*-gram frequencies are profoundly related to their online processing and processability. But also in other areas of corpus linguistics and computational linguistics, a lot of research involving *n*-grams has yielded many interesting results. For example, there is a wealth of studies that show that *n*-grams can be a good diagnostic or a good discriminatory tool in many corpus-linguistic and computational-linguistic domains, for example:

- lexical *n*-grams are used for multidimensional register classification (cf. Crossley & Louwerse 2007), the study of academic English (cf. Biber, Conrad & Cortes 2004 and Simpson-Vlach & Ellis 2010), the identification of junk/spam emails (Orasan & Krishnamurthy 2002), etc.;
- character *n*-grams are used for the identification of languages in web data (cf. Cavnar & Trenkle 1994), for spell-checking (cf. Memushaj & Sobh 2008), etc.;
- Solan et al. (2005) develop an unsupervised grammar induction algorithm that is ultimately based on *n*-gram frequencies and performs with a very high level of precision.[1]

In the light of the technological progress and the convergence of interests and findings from different angles, corpus linguists are therefore well advised and in a good position to explore this issue further. Note in this context that the study of *n*-grams is recent enough for the field not to have yet accepted standards on how to generate, explore, quantify, and study *n*-grams. In spite of the many fascinating findings mentioned above, there are certainly several areas where improvements are possible and worth exploring. In this paper we will explore various ways in which the study of *n*-grams may be refined or improved. The approach adopted in this paper is very different from all other *n*-gram studies we are aware of and can therefore be characterized best in opposition to what appears to be current practice.

First, there is the issue of "corpus-drivenness". More specifically, many studies using *n*-grams claim to use a 'corpus-driven' approach (cf. Tognini-Bonelli 2001) in that they retrieve *n*-grams from corpora without recourse to any particular theory and then use quantitative exploration to characterize and/or identify the *n*-gram patterns. However, it is not clear at all whether all of these studies are in fact as corpus-driven as they could be because most studies decide *a priori* on a particular *n*. Currently, $n = 4$ appears to be most fashionable. While this decision does of course not undermine the findings presented in such studies, it begs the question whether one or more different *n*'s would have yielded better results ("better" in the sense of "more revealing" or, more technically, "explaining more of the variability in the corpus"). We are currently aware of only one such study: Gries et al. (under revision) study different genres on the basis of *n*-grams with $1 \leq n \leq 5$. While this is in general a more data-/corpus-driven approach than studies that *a priori* settle for one and only one *n*, even this approach does not allow for the fact that it may actually be best not to focus on any one *n* but let the length of each *n*-gram emerge, as it were, from the data — which is what we will do in the present study.

Second, there is the issue of corpus granularity, which in essence is *also* a question of how data-driven a study actually is. More specifically, we refer to the fact that many studies using *n*-grams explore different corpus parts or different corpora on only one level of granularity. That is, these studies explore, say, academic writing and contrast it with other written data or with academic spoken discourse. This is risky because corpus data can be categorized on many different levels, not all of which are equally useful *a priori* or equally borne out empirically, which raises general methodological questions concerning corpus homogeneity/granularity (cf. Kilgarriff 2001, and in particular Gries 2006). With regard to *n*-grams, Gries et al. (under revision) again take at least one step in the right direction: not only do they study different *n*'s (with the above caveat, though), but they also test which and how many *n*-grams are most useful to reliably distinguish modes (spoken vs. written), genres, and sub-genres in the ICE-GB and the BNC Baby. In the present study, we will explore two different levels of granularity. More specifically, we will not only explore *n*-grams in speech versus in writing, but at the same time across different varieties, which brings us to the next point.

Third, many studies, very few of which we mentioned above, use *n*-grams to describe different registers, or genres, and have yielded many interesting findings.[2] In this study, however, we will explore to what degree the discriminatory power *n*-grams have exhibited with regard to genres is also obtained on the different (higher) level of resolution of different varieties of English. More specifically, we will follow up on our previous work on Asian components of the International Corpus of English (ICE) to which we will turn in Section 1.2 (cf. Mukherjee & Gries 2009).

Finally, nearly all studies involving *n*-grams are based on either raw frequencies of *n*-gram occurrence, or they are based on the collocational statistics of Mutual Information (*MI*). However, the former can be problematic because raw frequencies of occurrences can be rather misleading, which is why many corpus-based studies on co-occurrence phenomena use measures of collocational/collostructional attraction. It is only at first sight that this seems to strengthen the case for using *MI*: in virtually all existing studies of *n*-grams the application of *MI* is in essence incorrect. As far as can be seen, the standard software packages compute an *MI*-value for a 4-gram that compares the observed frequency of occurrence against an expected frequency computed on the assumption of complete independence, which is, as a matter of fact, almost never the case in natural language: for example, the probability of *of* two words after *in* is very much higher when the word immediately after *in* is *spite*. In addition, since these studies usually take a particular predetermined *n* as their starting point, the collocational measure is not used to identify varying length *n*-grams and, thus, unable to recognize that the right *n* may be 2 (to identify *shut up*), 3 (to identify *in spite of*), 4 (to identify *on the other hand*), 6 (to identify *at the end of the day*), etc. In the present study, we will therefore use a different approach that is not just based on raw frequencies of (co-)occurrence, but on a new and so far apparently underutilized measure of collocational attraction. This will be discussed further in Section 1.3.

1.2 Asian Englishes in the focus of corpus linguistics

As already mentioned above, in this paper we will study *n*-grams not with an eye to registers or genres, but to regional varieties of English. On the one hand, this will allow us to determine to what degree *n*-gram-based methods are also relevant to research into varieties, an approach that has not been undertaken so far. On the other hand, we have already shown in earlier work (Mukherjee & Gries 2009) that lexicogrammatical co-occurrence preferences of verbs and constructions can help to distinguish between different varieties of English. In the present study, we investigate whether the same is true for *n*-grams. By their very nature, *n*-grams are also located at the interface between lexis and syntax, which, according to Schneider (2007: 86), harbors "many of the characteristic innovations of PCEs [Post-colonial Englishes] […]: they concern the co-occurrence potential of certain words with other words or specific structures". Against this background, it is high time to include studies of *n*-grams in the growing body of lexicogrammatical research into differences between Englishes world-wide. In the present study, we will restrict ourselves to Asian Englishes as one of the most important groups of postcolonial New Englishes.

Let us first provide some background information on the current state of research into Asian Englishes. In the past few years, there has been a significant shift from research into individual Asian varieties of English to more integrated perspectives on the manifestations and realities of English in Asia. McArthur (2003), Kachru (2005) and Bolton (2008), for example, have introduced category labels such as "English as an Asian language", "Anglophone Asia" and "Asian Englishes", respectively, to capture the widespread use of English in many Asian countries both as a postcolonial link language in multilingual speech communities and as a pan-Asian communicative vehicle and a key to international communication. The focus of the present paper is on the former type of Asian Englishes, i.e. manifestations of English in postcolonial contexts in which English has been retained as an official, co-official or quasi-official language, continues to be routinely used in a wide range of communication situations and has developed into more or less localized "new" varieties of English.

There is a rich body of literature on the phonetic-phonological, lexical and grammatical features of many postcolonial Asian Englishes, ranging from feature checklists (e.g. Trudgill & Hannah 2002) and descriptive handbooks (e.g. Mesthrie 2008) to theory-driven models of World Englishes and Asian Englishes (e.g. Schneider 2007) and, most importantly, many descriptive studies of individual features of particular Asian Englishes (based on survey data, literary texts, individual examples, unsystematically collected datasets and, more recently, corpora). In the present paper, we will combine corpus-linguistic methodology with a descriptive-comparative approach to various postcolonial Asian Englishes, in particular English in Hong Kong, in India, and in Singapore. From a corpus-linguistic perspective, comparative studies of this kind are now much more easily possible than in the past because the International Corpus of English (ICE) includes representative corpora of the Asian Englishes under scrutiny (ICE-HK, ICE-IND and ICE-SIN) that are of the same size (1 million words each) and that have been designed according to the same principles, including the same amount of words from spoken texts (60%) and written texts (40%) from the 1990's and representing the same genres (cf. Greenbaum 1996). What is more, the three Asian Englishes share the same historical input variety, namely British English, for which ICE-GB, a fully parsed corpus, is available as a reference corpus (cf. Nelson et al. 2002).[3]

The reason why we chose Hong Kong English, Indian English and Singapore English as target varieties is that they represent different stages in the evolution of New Englishes, a process for which Schneider (2003, 2007) has recently suggested an innovative and ambitious model of variety-formation. One of the kernel ideas in Schneider's (2007) dynamic-evolutionary model is the assumption that New

Englishes have been — and are being — shaped according to a fundamentally uniform pattern world-wide: there is growing group-interaction between the indigenous people ('IDG strand') and the settlers, i.e. the colonists ('STL-strand'), across time, which leads to a more and more integrated, new and hybrid identity-construction, which in turn manifests itself in a new variety of English marked by 'structural nativization', i.e. "the emergence of locally characteristic linguistic patterns" (Schneider 2007: 5f.). The evolution of New Englishes is described as a succession of five characteristic stages by Schneider (2003, 2007):

- Phase I–'Foundation': In this initial phase, the English language is transported to a new (colonial) territory.
- Phase II–'Exonormative stabilization': There is a growing number of English settlers/speakers in the new territory, but the language standards and norms are still determined by the input variety and are, thus, usually oriented towards British English.
- Phase III–'Nativization': The English language becomes an integral part of the local linguistic repertoire as there is a steady increase in the number of competent bilingual L2 speakers of English from the indigenous population.
- Phase IV–'Endonormative stabilization': After Independence, English may be retained as a/an (co-)official language and a medium of communication for a more or less wide range of intra-national contexts (e.g. administration and the press, academia and education); in this phase a new variety of English emerges with generally accepted local standards and norms.
- Phase V–'Differentiation': Once a New English variety has become endonormatively stabilized, it may develop a wide range of regional and social dialects.

We have argued elsewhere that the varieties of English in Hong Kong, India and Singapore represent distinctly different phases in the evolutionary process (cf. Mukherjee & Gries 2009: 31ff.): while Hong Kong English can be mapped onto phases II and III, Indian English displays features of phases III and IV, and Singapore English is a prototypical example of an advanced phase IV variety. The comparison of ICE corpora representing the three Asian Englishes and British English is promising and insightful as it is possible now to trace potential correlations between the evolutionary stage of a variety and its degree of structural nativization at the morphological, lexical, lexicogrammatical and syntactic level.

In spite of a growing interest amongst a number of linguists, the lexis-grammar interface is still largely a blind spot in research into many postcolonial varieties of English. This has to do with the fact that at the lexicogrammatical level, e.g. with regard to collocations and verb-complementational patterns, differences between varieties of English are usually not categorial but quantitative in nature, so that large and representative corpora are needed to identify different trends and preferences across varieties of English. Before the advent of ICE, balanced (let alone, comparable) corpora had not been available for most New Englishes. The ICE family of corpora enables us to describe lexicogrammatical differences between varieties on an empirically sound basis. This is particularly relevant also because lexicogrammatical differences, e.g. the preferred use of the monotransitive pattern with the verb *give* in Indian English as opposed to the preferred use of the ditransitive pattern with *give* in British English (cf. Mukherjee & Hoffmann 2006), "operate way below the level of linguistic awareness: without quantitative methodology no observer would have expected such differences to exist" (Schneider 2007: 87). That is, while nativization at the lexical level (e.g. borrowings from local languages) and the syntactic level (e.g. different use of the article system and tense usage) are often more or less active choices and/or shaped by interference effects, structural nativization at the lexis-grammar interface is much more subtle, opaque and gradient — and, in a sense, more inherent to the structural characteristics of a variety.[4]

Against this background, in our previous work we investigated collostructions, i.e. verb-construction associations as defined by Stefanowitsch & Gries (2003), across ICE-HK, ICE-IND, ICE-SIN and ICE-GB (cf. Mukherjee & Gries 2009). Specifically, we analyzed four different groups of verbs: 15 verbs attracted to the ditransitive construction in ICE-GB, 14 verbs attracted to the intransitive construction in ICE-GB, 15 verbs attracted to the monotransitive construction in ICE-GB, and 15 verbs with no constructional preferences ('neutral verbs') in ICE-GB. All instances of these 59 verbs in the four corpora in any of the following three constructions were taken into consideration: (a) the ditransitive construction, (b) the intransitive construction, (c) the monotransitive construction. In a first step, a proportional sample of the 59 verbs in the three constructions in the three Asian English varieties was constructed ($n = 11{,}487$). In a second step, all 11,487 instances were handcoded as ditransitive, intransitive, monotransitive or as other/non-canonical. By conducting a Multiple Distinctive Collexeme Analysis (MDCA), we then calculated for the pool of all instances handcoded as ditransitive, intransitive and monotransitive whether and to what extent the co-occurrence of a given verb and a construction was higher (+) or lower (–) than expected. The analysis resulted in a matrix of constructional preferences and dispreferences for each of the 59 verbs across the four corpora. One of the interesting results of this study was that various verbs that prefer the ditransitive construction in British English do not prefer the ditransitive construction in one of the Asian Englishes (e.g. *convince* in Hong Kong English, *cost* in Indian English and *lend* in Singapore English), and that in all of these cases, the verbs at hand prefer the monotransitive construction in the Asian English variety. The most important overarching finding, however, was that a clear correlation between the matrix of constructional preferences and dispreferences of the 59 verbs in an Asian variety on the one hand

and its evolutionary stage on the other emerged from the data: "the more advanced a New English variety is in its evolution, the more dissimilar it is to BrE at the level of collostructions" (Mukherjee & Gries 2009: 47f.).

Our previous work, therefore, confirmed the evolutionary model posited by Schneider (2007) at the level of verb-construction associations in three Asian Englishes. It also showed that the ICE corpus set we used, consisting of three Asian Englishes in clearly different evolutionary stages and the present-day form of the shared historical input variety, is a useful dataset for quantitative intervarietal comparisons at the lexis-grammar interface. In the light of our previous work, it seemed very promising to use the same dataset of ICE components for our new corpus-driven approach to *n*-grams as introduced in Section 1.1.

1.3 *N*-grams and the measurement of their attraction

As sketched out above, most studies involving *n*-grams use raw frequencies of occurrence or *MI*-values as their main diagnostic. We also already mentioned that this procedure is somewhat problematic, given how *MI*-values are typically computed. However, there is yet another problem that is in fact far more pertinent and applies to the 30 or so measures of collocational strength that have been proposed (cf. Wiechmann 2008 for a comprehensive overview). This problem has to do with the fact that nearly all collocational measures are exclusively based on token frequencies and do not take into account type frequencies. In essence, measures of collocational strength are generally based on 2×2 co-occurrence tables of the type represented in Table 1.

Table 1. Schematic lexical co-occurrence table

	word *y*	**not word *y***	**Totals**
word *x*	*a*	*b*	*a+b*
not word *x*	*c*	*d*	*c+d*
Totals	*a+c*	*b+d*	*a+b+c+d*

Cell frequency *a* represents the same thing in all measures, namely the frequency of the co-occurrence of *x* and *y*. However, for all measures, frequencies *b* and *c* are the token frequencies of *x* where *y* is not given and *y* where *x* is not given, respectively, but this means that the type frequency of cells *b* and *c* is not figured into the measure(s). That is, if *b* = 900, i.e. there are 900 occurrences of *x* without *y*, then all regular measures use the number 900 for the subsequent computation, regardless of whether these 900 tokens consist of 900 different types or of 2 different types. This is potentially problematic since an important dimension of variation in the data is disregarded even though it is well known that type frequencies are in fact very important in a variety of areas:

- (constructional) acquisition in first language acquisition where children in fact develop syntactic knowledge out of *n*-grams that allow their slots to be filled with different degrees of lexical flexibility (cf. Goldberg 2006: Ch. 5);
- as determinants of language variation and change (cf. Hopper & Traugott 2003);
- as a correlate of measures of (morphological) productivity (cf. Baayen 2001).

A potentially better measure of collocational attraction would therefore take type frequencies into consideration, but as even a comprehensive search will show, there have been very few suggestions in this regard in previous research. A study which uses an approach that has apparently never been replicated is Kita et al. (1994). They use a bottom-up approach to *n*-grams. An easy example to explain their so-called 'cost criterion' is the identification of the *n*-gram *in spite of*, namely by recognizing that, informally speaking, there is so little variation after *in spite* that it makes more sense to consider *in spite of* as a unit in the first place.

In the present study, we are going to employ a similar approach based on a new measure of collocational strength as developed by Daudaravičius & Marcinkevičienė (2004). This measure, gravity *G*, takes type frequencies into consideration, as is indicated in Equation (1).

$$(1)\quad GravityG(word_1,word_2) = \log\left(\frac{freq(word_1,word_2)\cdot typefreqafterword_1}{freqword_1}\right) + \log\left(\frac{freq(word_1,word_2)\cdot typefreqbeforeword_2}{freqword_2}\right)$$

The meaning of this formula is not immediately obvious; given the importance of this formula for what follows, let us briefly exemplify, and illustrate graphically, how this measure works. As (1) shows, the computation of *G* involves five different frequencies:

- the co-occurrence frequency of the two adjacent words: freq($word_1$, $word_2$);
- the frequency of the first word: *freq* $word_1$;
- the frequency of the second word: *freq* $word_2$;
- the frequency of types after the first word: *type freq after* $word_1$;
- the frequency of types before the second word: *type freq before* $word_2$.

Each of these frequencies can influence the size of *G*, which is why the measure can be understood best by considering the effect of any one of these frequencies when all others are held constant. Consider Figure 1.

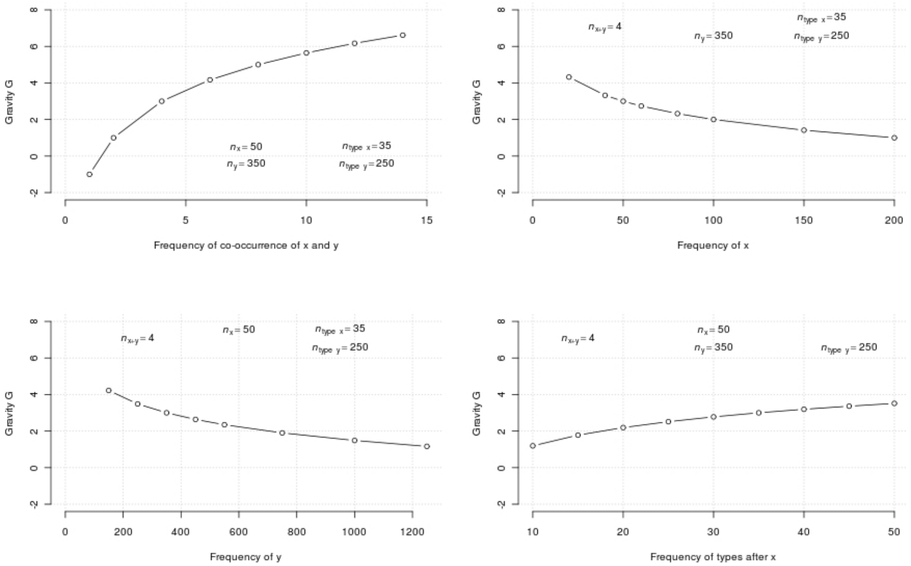

Figure 1. The behavior of *G* depending on its components

Each panel in Figure 1 represents how *G* (on the *y*-axis) changes when only one of the frequencies of the formula in (1) is changed, namely the co-occurrence frequency (on the *x*-axis); all the frequencies held constant are plotted into the coordinate system. The upper left panel shows that when the frequencies of both words and the frequencies of types before and after them are held constant, then *G* increases as the observed frequency of co-occurrence increases. This is, of course, intended because if the observed frequency of a collocation increases although the frequencies of its component words do not and although there are not fewer different types around the component words, then this collocation's *G*-value becomes larger.

Analogously, the upper right panel represents how *G* (on the *y*-axis) changes when another one of the frequencies of the formula in (1) is changed, namely the frequency of the first word (on the *x*-axis); all other frequencies are again held constant and are plotted into the coordinate system: when the frequency of the first word becomes larger, but everything else stays the same — including the co-occurrence frequency — then *G* decreases for a simple reason: when one word of a bigram becomes more frequent, more co-occurrences are expected even by chance, and if that number does not increase, then the collocation is weaker. The same logic applies to the lower left panel, which shows how *G* reacts in the same way to an increase of the second word in the collocation.

Finally, the lower right panel illustrates how *G* reacts when the type frequency after the first word increases, with all other frequencies remaining the same (as indicated in the plot). When the type frequency after the first word increases, but the co-occurrence frequency stays the same, then this means that the collocation can be observed the same number of times even if there are more different types, which would potentially lower the collocation's expected frequency, and hence *G* increases.

There are as yet very few studies that use *G*-values of this kind. In fact, the only one we are aware of is Gries (2010a), which studies how well gravity-based 2-grams can distinguish between modes, registers, and sub-registers in the BNC Baby and ICE-GB. Interestingly, he finds that cluster analyses based on the *G*-values of bigrams not only reproduce — almost ideally — the structures of the corpora even down to the level of clustering sub-registers into meaningful groups, but also outperform *t*-scores. Given *G*'s overall behavior, in particular its responsiveness to type frequencies, and the above-mentioned importance of type frequencies, *G* appears to be a particularly important measure to explore since it is well known that type frequencies are in fact very important in a variety of areas.

While the gravity calculation in (1) only applies to 2-grams, Daudaravičius & Marcinkevičienė (2004) took the idea further. They proposed to include *n*-grams with $n > 2$ by identifying *n*-grams consisting of successive 2-grams whose *G*-value exceeds the threshold value of 5.5. Figure 2 exemplifies one sentence.

In the present study, we will apply gravity calculations of this kind to the description of lexicogrammatical differences between varieties of English, focusing

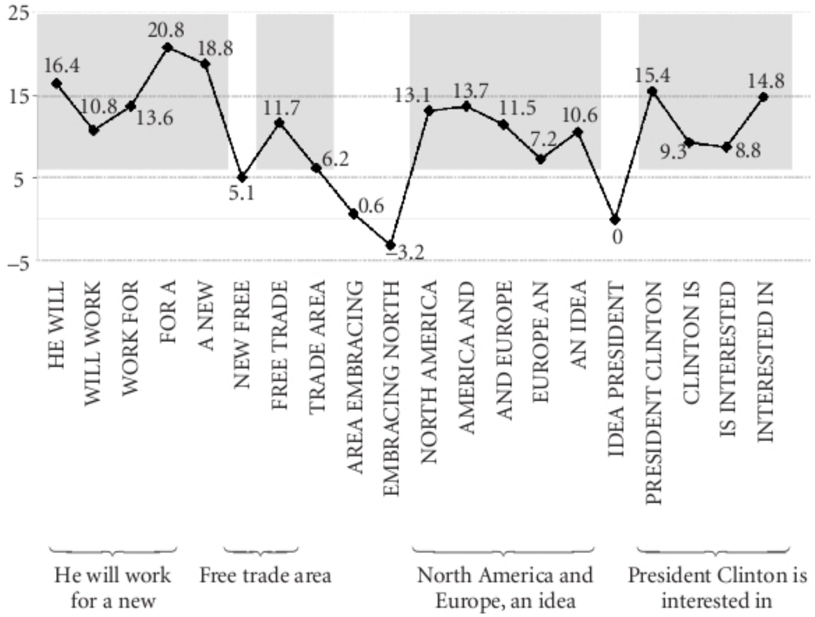

Figure 2. Using *G*-values to identify "collocational chains" (from Daudaravičius & Marcinkevičiené 2004: 334)

on Asian Englishes and using comparable ICE components. The remainder of the paper is structured as follows. Section 2 discusses a first case study, an exploration of *n*-gram patterns in Asian Englishes. In Section 2.1 and 2.2, we discuss our methodology in detail, as it exhibits several crucial differences compared to most other *n*-gram work, while Section 2.3 discusses the results and their methodological implications. Section 3 is concerned with a small follow-up exploration of these results. Section 3.1 will discuss the methodology of the follow-up study, while Section 3.2 provides the results. In Section 4, we will offer some conclusions both at the descriptive and at the methodological level.

2. Case study 1: *N*-grams in English varieties based on gravity values

2.1 Methods–Part 1: The study of 2-grams

In the present work, we study *n*-grams in eight corpus parts, referring to speech and writing in four varieties:

- ICE-GB: spoken vs. written;
- ICE-HK: spoken vs. written;
- ICE-IND: spoken vs. written;
- ICE-SIN: spoken vs. written.[5]

For each corpus part, we identified *n*-grams using the following algorithmic procedure: First, we extracted all words and then all 2-grams case-insensitively within each sentence of the corpus. This means that 2-grams consisting of the last word of one sentence and the first word of the following sentence were not included. For ICE-GB, words were defined as those character strings that were between curly brackets in the annotated files and that contained at least one letter (including spaces, hyphens, and apostrophes); for the other varieties, words were defined using the regular expression "[^-a-z\\']+".[6] For all varieties, we adopted the corpus compilers' definition of sentences as indicated in the corpus annotation. Second, for each of the 2-grams thus obtained, we computed the gravity value *G*. By way of exemplification, consider Table 2 for these interim results for three sentences from the spoken part of ICE-GB (Sentence 1: *Excuse me, I've got to do what I did last time*. Sentence 2: *I hate this*. Sentence 3: *I've got to get this out*.).

By way of example, (2) shows how *G* is computed for the first 2-gram, *excuse me*:

$$(2)\quad G_{\text{excuse me}} = \log_2\left(\frac{12 \cdot 10}{25}\right) + \log_2\left(\frac{12 \cdot 272}{1176}\right) = 2.263034 + 1.472753 = 3.735787$$

Table 2. 2-grams and their G-values for three randomly-chosen consecutive sentences

Sentence	2-gram	$n_{\text{2-gram}}$	$n_{\text{word 1}}$	$n_{\text{word 2}}$	n_{types}	n_{types}	G
					after w1	before w2	
16	*excuse me*	12	25	1176	10	272	3.74
16	*me i*	18	1176	15475	234	1148	2.26
16	***i 've***	**861**	**15475**	**2113**	**705**	**42**	**9.39**
16	***ve got***	**661**	**2113**	**1591**	**259**	**118**	**11.96**
16	***got to***	**242**	**1591**	**15505**	**284**	**2593**	**10.77**
16	***to do***	**591**	**15505**	**2379**	**2179**	**264**	**12.41**
16	*do what*	22	2379	3293	284	446	2.97
16	***what i***	**311**	**3293**	**15475**	**377**	**1148**	**9.68**
16	***i did***	**125**	**15475**	**878**	**705**	**195**	**7.3**
16	*did last*	4	878	504	162	175	0.04
16	*last time*	28	504	1114	164	172	5.3
17	*i hate*	3	15475	18	705	9	−2.29
17	*hate this*	1	18	3939	13	674	−3.02
18	***i 've***	**861**	**15475**	**2113**	**705**	**42**	**9.39**
18	***ve got***	**661**	**2113**	**1591**	**259**	**118**	**11.96**
18	***got to***	**242**	**1591**	**15505**	**284**	**2593**	**10.77**
18	***to get***	**351**	**15505**	**1161**	**2179**	**115**	**10.74**
18	*get this*	8	1161	3939	249	674	1.23
18	*this out*	10	3939	1472	968	483	3.01

Based on these G-values, we ran four different hierarchical agglomerative cluster analyses on four different data matrices.[7] The first two of these disregarded Daudaravičius & Marcinkevičienė's (2004) threshold value of 5.5 for *G*-values as well as their recommendation to only compute *G*-values for 2-grams consisting of words whose joint frequency exceeds 10; the second two did not:

- for the first cluster analysis, we generated a matrix with all 12,531 2-grams that occurred in each of the eight corpus parts (four varieties crossed with two modes) in the rows, with the eight corpus parts in the columns, and the *G*-value for each such 2-gram in each corpus part in the cells; consider Table 3 for the first six rows of this matrix;
- for the second cluster analysis, we generated a matrix with all 47,524 2-grams that occurred in each of the four variety corpora in the rows, with the four variety corpora in the columns, and the *G*-value for each such 2-gram in each corpus part in the cells;

- for the third and fourth cluster analyses, we trimmed down the above matrices of *G*- values in corpus parts to one containing only 491 2-grams that met both conditions, and we ran two cluster analyses on these, too.

Table 3. The input table for the first cluster analysis (which clustered corpus parts according to *n*-gram collocational strengths measured in G)

2-gram	GB-s	GB-w	HK-s	HK-w	IND-s	IND-w	SIN-s	SIN-w
a a	11.474	−0.872	11.619	−1.007	7.305	3.329	9.698	0.106
a an	6.334	−0.046	5.259	−3.379	4.288	2.4	2.434	−3.234
a and	7.943	5.965	8.282	2.429	5.96	5.929	4.645	5.671
a b	2.228	5.49	2.224	−3.614	4.524	7.396	5.376	2.611
a bad	6.225	4.304	2.812	4.471	4.434	2.711	4.544	2.839
a balance	1.052	2.099	0.576	4.336	−0.973	0.693	3.462	3.266
…	…	…	…	…	…	…	…	…

In Section 2.3.1, we will discuss the results for the 2-grams.

2.2 Methods–Part 2: The study of larger *n*-grams

In addition to this simplifying evaluation with 2-grams only, we also extended Daudaravičius & Marcinkevičienė's (2004) approach in order to be able to include *n*-grams with $n > 2$. As mentioned above, in their original paper, they proposed to extract what they refer to as 'collocational chains' by identifying chains of 2-grams that exceed their proposed threshold value of 5.5 (recall Figure 2), and these *n*-grams are highlighted in bold type in Table 2. However, one potential problem with this approach is that it does not base the length of their collocational chains on the number of words that have already been added to them with different *G*-values. For example, one 3-gram may consist of two 2-grams with very high *G*-values (e.g. 20 and 22) while another 3-gram may consist of one 2-gram with a very high *G*-value (e.g. 20) and one 2-gram with a *G*-value that is just about large enough to exceed the threshold of 5.5. Crucially, the proposed approach would treat these two hypothetical 3-grams in very much the same way. That is, this approach is greedy in that it looks for the longest possible *n*-gram whose 2-gram *G*-values exceed 5.5 but it does not contain a cut-off that would, for instance, let it decide that although all the 2-grams of *i've got to do* have a *G*-value greater than 5.5, the empirically better motivated *n*-gram may be *i've got to*.

In order to address this potential complication, we also computed each *n*-gram's mean *G* (cf. again Table 2 for the *G*-values that enter into the computations for these examples):

- *i've got to do*: $mean_G = 11.13$ (length of *n*-gram $l = 5$);
- *what i did*: $mean_G = 8.49$ (length of *n*-gram $l = 3$);
- *i've got to get*: $mean_G = 10.72$ (length of *n*-gram $l = 5$).

In a final step, for each *n*-gram *N* of length *l* and (mean) $G \geq 5.5$ we then tested whether there is another *n*-gram that

- contains the first *n*-gram *N*;
- has a length $l+1$;
- has a higher $mean_G$-value.

This test yielded two possible outcomes. If we found that there was no such longer *n*-gram, we retained the shorter *n*-gram *N* because it could not be seen as a part of a larger *n*-gram with a higher average degree of cohesion; or if we found that there was a larger *n*-gram, the longer *n*-gram(s) was/were retained and *N* was discarded. Consider Table 4 for one example of each possible outcome (again from the spoken part of the ICE-GB). As an example of the former, consider the 2-gram *for the*: it has a *G*-value of 13.17, and none of the 3-grams containing *for the* has a $mean_G$-value greater than that, which is why *for the* was retained as a 2-gram. As an example of the latter, consider the 2-gram *might have*: it has a *G*-value of 6.84, but the 3-grams *might have a* and *might have had* have a $mean_G$-value greater than that, which is why *might have* was not retained as a 2-gram.

As a result of applying this multi-step procedure to each of the eight corpus parts, we obtained a list of *n*-grams of different lengths and their (mean) *G*-values

Table 4. The identification of *n*-grams with $n > 2$: two examples

***N*-gram**	***N*-gram length**	$mean_G$
for the	2	13.17
might have	2	6.84
as as	2	6.98
for the moment	3	11.3
pay for the	3	9.87
for the work	3	10.54
might have a	3	9.33
might have had	3	6.88
it might have	3	6.31
you might have	3	6.55
might have done	3	6.48
…	…	…

for each corpus part. Again, we merged the lists from the different corpus parts into tables of exactly the same kind as exemplified in Table 3 and applied a hierarchical agglomerative cluster analysis to identify what (if any) structure the *n*-gram patterns exhibit across the four varieties and in their spoken and written parts. The results of all these cluster analyses of larger *n*-grams will be discussed in Section 2.3.2.

2.3 Results

2.3.1 *Results for 2-grams*

The results of the first two explorations are shown in Figure 3, which indicates clearly that the *G*-values of the 2-grams clearly distinguish speech from writing across all four varieties. However, the cline of the evolutionary stages of the four varieties (Singapore English most advanced, Hong Kong English least advanced) is not replicated by the dendrograms: rather, Hong Kong and Singapore English are most similar to each other, followed by British English and then Indian English.

The results of the second two cluster analyses in which both recommendations by Daudaravičius & Marcinkevičienė (2004) are followed (and in which we only included *G*-values greater than or equal to 5.5 and only 2-grams whose words had a combined frequency of more than 10), the picture is very similar. As Figure 4 shows, the *G*-values again clearly distinguish speech from writing, but there is no clustering of the varieties that is compatible with the evolutionary model suggested by Schneider (2007).

2.3.2 *Results for n-grams*

The final set of cluster analyses is based on variable-length *n*-grams as defined by the above algorithm. Again, the results, as shown in Figure 5, are similar: there is

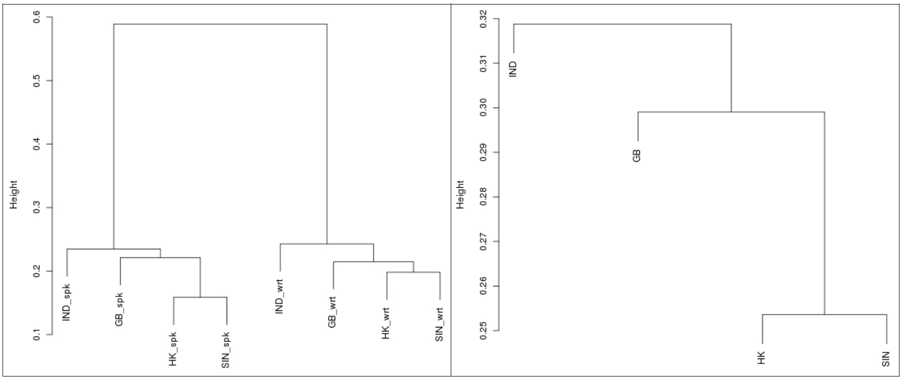

Figure 3. Dendrograms from first exploratory cluster analyses of 2-gram gravities (disregarding *G*-values and combined word frequencies)

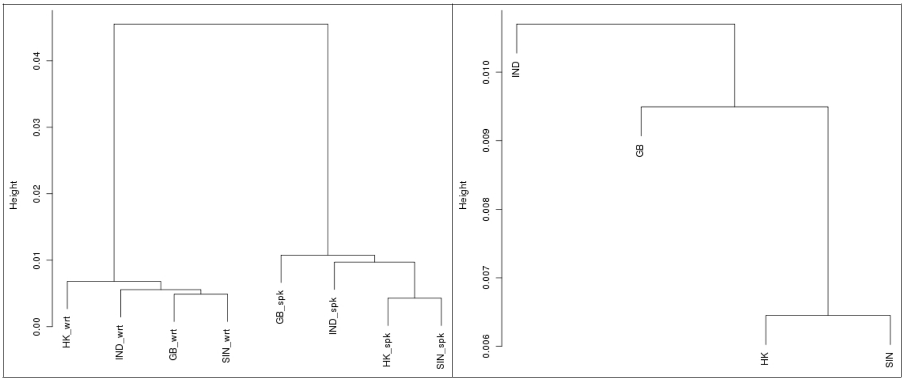

Figure 4. Dendrograms from exploratory cluster analyses of 2-gram gravities (with *G*-values ≥ 5.5 and combined word frequencies > 10)

a neat clustering of speech versus writing across varieties, but no replication of the evolutionary stages as suggested by Schneider (2007).

2.3.3 *Implications*

Figures 3 to 5 provide two interesting findings. First, at the level of *n*-grams the differences between speech and writing figure prominently in British English and all three Asian Englishes. This result indicates that in New Englishes, too, there is a clear differentiation in usage between the two media and that, at least at the level of *n*-grams, speech is not a hyper-formal (or "bookish", cf. Kachru 1983: 39) imitation of writing — Asian speakers of English as a second language clearly distinguish between the two media in their *n*-gram usage. Second, the evolutionary stages of the three Asian Englishes are not replicated by the cluster analyses. Given that the collostructional findings from the same corpora on which Mukherjee & Gries (2009) report are perfectly in line with Schneider's (2007) evolutionary model, the question

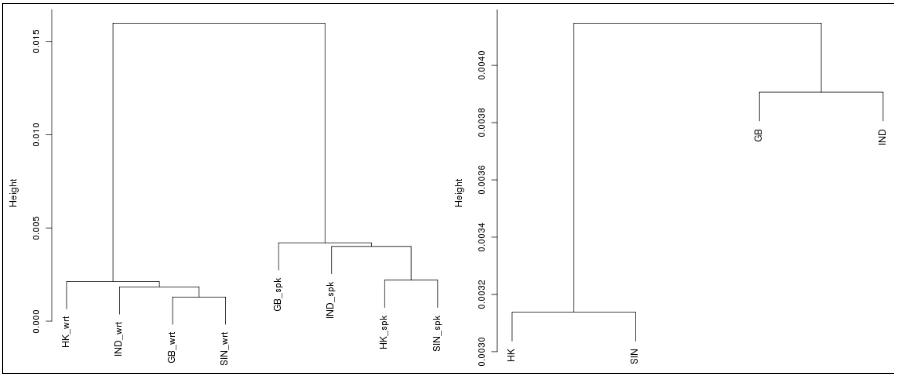

Figure 5. Dendrograms from exploratory cluster analyses of *n*-gram gravities

arises why the sequence of stages as represented by ICE-HK, ICE-IND and ICE-SIN cannot be found at the level of *n*-grams. There are various potential answers:

- The gravity measure may be flawed — but the logic underlying it appears very well-founded, and the spoken-written distinction is identified correctly and reliably in each analysis.
- The evolutionary model is inaccurate — but the model is based on a substantial body of work and it has been supported by a wide range of case studies (cf. Schneider 2007: 113ff.) and previous work on Asian Englishes based on ICE corpora (cf. Mukherjee & Gries 2009).
- The evolutionary model does not apply to lexis/*n*-grams in the sense that different evolutionary stages are less likely to be reflected in lexical differences because lexis-related statistics are topic-dependent and volatile. Could it be that something more robust is needed?

We tend towards the third answer. In spite of the fact that *n*-grams have been a useful approach towards the identification and characterization of genres, they may be too fine-grained and too volatile a method to indicate evolutionary differences between different varieties in corpora. There are various reasons why this may indeed be the case. First, just like all other lexical statistics they are highly sensitive towards topics or topical domains, which is why they are useful for the identification of genres and why, for example, Leech & Fallon (1992) explain some of their findings from corpus comparisons with regard to topics that were relevant at the time of compilation of the Brown corpus (e.g. the Cuba Missile crisis). Second, the corpora studied here represent rather heterogeneous datasets after all since they have intentionally been sampled to contain data from many different genres.

Even if *n*-grams are too volatile *per se* for our purposes, they may, however, still be useful at a coarser level of granularity. To illustrate this point, consider Figure 6. One could argue that Mukherjee & Gries (2009) showed that the rightmost, coarsest level of analysis resulted in results compatible with Schneider's (2007) evolutionary model, while the present section showed that the leftmost, finest level of analysis did not. The obvious question that follows from Figure 6 is: What about the intermediate level?

To us, it certainly seems interesting — and necessary — to explore the intermediate level of analysis in Figure 6, that is, to address the question of how much words "like" to be in *n*-grams in the first place, i.e. how "sticky" they are in general.

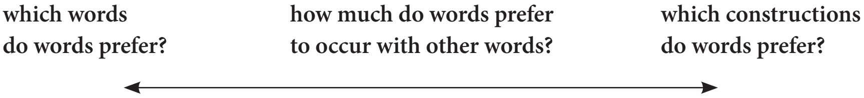
which words do words prefer?
how much do words prefer to occur with other words?
which constructions do words prefer?

Figure 6. Levels of resolution in the analysis of word-specific preferences

More specifically, for each of the eight corpus parts, we obtained a large number of *n*-grams of different sizes. The intermediate perspective in Figure 6 then raises the question of which (kinds of) words are "sticky" in the sense that they are strongly over-represented in *n*-grams compared to their frequency in the corpus part at hand. The lead question for this kind of analysis could be formulated as follows: What are the words that prefer to occur in "significant" *n*-grams over being used individually, and do these words differ across the eight corpus parts, i.e. across the four varieties and across speech and writing? To this question we will turn in the following section.

3. Case study 2: Lexical stickiness

3.1 Methodology

In order to explore the notion of stickiness, we took all the *n*-grams that our extension of the gravity measure returned, split them up into their component words and then explored two approaches that are conceptually similar to the keywords approach that has become so popular in corpus linguistics. The first one is relatively simple and is based on the frequencies with which component words are observed in *n*-grams. We created a table with these component words in the rows, the eight corpus parts in the columns, and the frequencies of the component words in *n*-grams in the cells. As an example, consider Table 5, which shows for the six most frequent words how often they were part of an *n*-gram in each of the eight corpus parts.

These frequencies were then *z*-standardized column-wise and fed into a cluster analysis (using the same parameters as before).

The second approach is also based on the frequencies with which component words are observed in *n*-grams, but is more complex and involves three steps.

Table 5. Computing a stickiness value for words (approach 1)

Word	GB-s	GB-w	HK-s	HK-w	IND-s	IND-w	SIN-s	SIN-w
the	6810	3948	11751	4337	7201	3630	6866	3536
to	4034	1884	7707	2037	3803	1468	4806	1642
of	3029	1764	3766	1800	2868	1870	2480	1632
i	3532	293	8175	525	2254	146	3263	253
that	3886	774	5346	692	3212	662	2963	742
and	3002	957	5401	955	2925	738	3088	830
…	…	…	…	…	…	…	…	…

Table 6. Computing a stickiness value for words (approach 2, based on ICE-GB spoken)

Word	In *n*-grams	In corpus part		% in *n*-grams	% in corpus part		Ratio
a	3377	13475		0.02986975	0.02069105		1.44
able	98	213		0.0008754012	0.0003285756		2.66
able-bodied	0	21	→	8.842437e-06	3.37788e-05	→	0.26
…	…	…		…	…		…
analysis	1	53		1.768487e-05	8.291161e-05		0.21
…	…	…		…	…		…

First, for each of the eight corpus parts, we again determined the frequencies of the component words in *n*-grams (as represented in Table 5), but also the overall frequencies of the component words in the corpus part at hand. To each of these frequencies we added 1 (to avoid problems due to many zeroes). To illustrate, consider the left panel of Table 6, which shows that in the spoken part of ICE-GB the word *able* occurs 98 times in *n*-grams and 213 times altogether. Second, we converted these frequencies — for *able*, 98 and 213 — into column percentages by dividing each frequency by the total of that column: the frequencies of *able* in *n*-grams (98) and in general (213) amount to 0.0875% and 0.0329% respectively; cf. the middle panel of Table 6. Third, for each corpus part, we subsequently divided the values in the "% in *n*-grams" column by those in the "% in corpus part" column, yielding the ratio shown in the right panel of Table 6. As Table 6 shows, *able* is 2.66 times more frequently part of an *n*-gram than its overall frequency in the corpus would suggest.

For a subsequent cluster analysis, we merged all of the corpus part-specific results into one table — with all words in *n*-grams in the rows, all corpus parts in the columns, and the ratio values from Table 6 in the cells, which was then again analyzed cluster-analytically.

3.2 Results

The results are interesting because they reveal how the stickiness of words interacts with the mode when it comes to distinguishing the four varieties. Consider the left and the right panel of Figure 7 for the results of the two approaches sketched out in Section 3.1. In both dendrograms the written genres are very similar to each other and there is virtually no discernible structure. The spoken genres, on the other hand, are different. In the first, simpler, approach they are quite different from the written data and exhibit a structure that is compatible with the evolutionary model: the Indian and the Singaporean data cluster together, as do the British and Hong Kong data. In the second, more complex approach, the Indian and the Singaporean data cluster together as well, as do the British and Hong Kong data, but the Indian and the Singaporean spoken data are more similar to the overall cluster of written data than to British and Hong Kong spoken data.

The findings shown in Figure 7 thus reveal that while speech and writing are significantly different at the level of *n*-grams across all four varieties, there are two interesting variational-linguistic nuances emerging from the data that may be explained by taking into account developmental features of the varieties at hand. First, the fact that Hong Kong English and British English speech cluster together may be explained by the low degree of nativization of Hong Kong English so that exonormative standards and norms (from British English) exert a much greater influence on Hong Kong speakers' language use in speech. This is also in line with the widely held view that the rare use of English for intraethnic communication in Hong Kong makes Hong Kong English "more like a foreign than a second language" (Li 2009: 74). This rather unstable and non-indiginized status of English, the future of which has become even shakier after the 1997 handover (cf. Li 1999), correlates with a comparatively low level of acceptance of local norms on the part of speakers of English in Hong Kong and a continuously strong exonormative orientation towards native standards set by the historical input variety.

Second, the fact that Indian English and Singapore English speech cluster together and are more similar to the written mode in general may indicate what Mesthrie & Bhatt (2008: 114) refer to as "'register shift', which clearly reflects the influence of written norms upon speech" in advanced and endonormatively stabilized varieties of English. Kachru & Smith (2008: 139f.), for example, discuss systematic differences in style between Indian English and other (native) varieties of English, based on cultural differences, which lead to a much more formal style in comparable contexts and situations in Indian English than in, say, British English. A recent large-scale corpus-based cross-varietal study by Xiao (2009), which reports on the

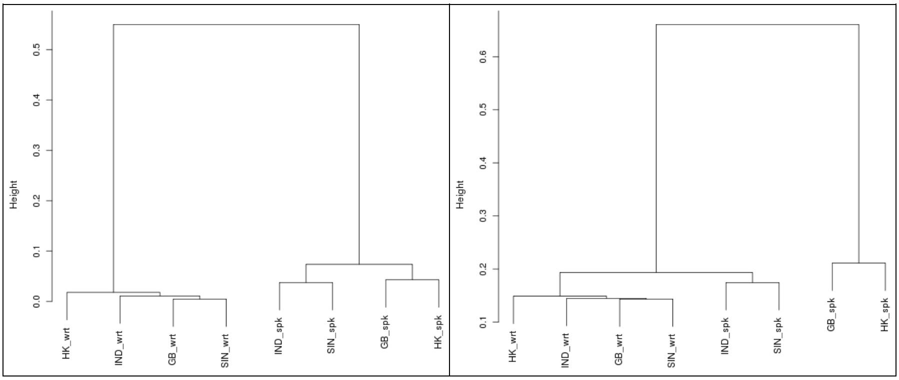

Figure 7. Dendrograms from clustering words according to their stickiness in *n*-grams

results of a multidimensional/multifactorial analysis of a large number of lexical and grammatical features (excluding, however, truly lexicogrammatical features such as *n*-grams) of various ICE components, corroborates the general observation that Indian English is in general more oriented towards written norms even in spoken language: "Indian English displays the lowest score for Factor 1 in nearly all registers, meaning that it is less interactive but more elaborate" (Xiao 2009: 442f.). Our cluster analyses indicate a similar trend towards written norms in Singapore English at the level of *n*-grams. Against this background, our findings thus tie in well with recent models of World Englishes and descriptions of Asian Englishes, but add new lexicogrammatical insights into Asian Englishes to the picture.

4. Concluding remarks

Our study of the characteristics of three Asian Englishes involved several methodological innovations:

- It involved one of the first attempts at validating a new collocational measure, lexical gravity *G*, that has the highly attractive feature of taking type frequencies within collocational slots into consideration.
- At the same time, we have suggested an extension of this measure that can determine the most relevant lengths of *n*-grams in a bottom-up and iterative way that is computationally easy and not particularly demanding. In this regard, at least, the present study is more corpus-driven than many so-called corpus-driven studies, since we let a collocational *n*-gram measure decide on the most appropriate *n*-gram length and since we use a bottom-up clustering method to seek structure in the *n*-gram data.
- We introduced a new measure to determine the degree to which a word prefers to be used in *n*-grams or patterns rather than on its own: lexical stickiness.

At the descriptive level, our findings show clear and identifiable differences between speech and writing in all four varieties; this differentiation, thus, seems to play a prominent role in all the evolutionary stages that are represented by the four varieties of English that we looked at. However, while Mukherjee & Gries (2009) found a clear alignment of the evolutionary stage of Hong Kong English, Indian English and Singapore English on the one hand and their lexicogrammatical patterning at the level of collostructional (dis-)preferences on the other, we could not replicate such a neat variety-to-form mapping for *n*-grams. It seems to us that *n*-grams are just too fine-grained and volatile a measure for this particular purpose: given their strong topic-dependence, they are often suitable for genre distinctions, but not for general variety distinctions *across genres*.

On the other hand, a more coarse-grained measure involving the notion of how much words like to be part of *n*-grams yielded results that are compatible with the postulated evolutionary stages for the spoken registers: the Hong Kong data cluster with the British data representing the input variety, and the Indian and Singaporean data cluster together as well and are closer to the written data, indicating a potential 'register shift' (cf. Mesthrie & Bhatt 2008). The written data in the four varieties are very similar to each other; there is hardly any clearly identifiable cluster structure. The homogeneity of the written data in our corpus set ties in with a general observation about differences between varieties of English, namely a tendency towards "convergence in writing, divergence in speech" (Mair 2007: 84).

In a wider setting, our observations and conceptual considerations raise general questions about the appropriate level of descriptive granularity at which evolutionary stages of the development of New Englishes manifest themselves. It seems to us that the grouping of New Englishes into evolutionary stages is a categorization at a very high level of abstraction; against this background it is obvious that a neat alignment of evolutionary stages on the one hand and linguistic features, their frequencies and distributions on the other can only be found at the level of rather abstract linguistic configurations based on a wide range of linguistic forms. In Mukherjee & Gries (2009), we looked at verb-construction associations, which refer to comparatively abstract co-occurrence patterns, across 59 verbs and 3 constructions, i.e. a relatively large number of linguistic forms — it is thus not surprising that we were able to detect a clear alignment of evolutionary stages and variety-specific patternings of linguistic forms. At the level of *n*-grams, however, there is a much stronger lexical bias (also influenced by individual topics): word-word associations as such do not — and cannot be expected to — mirror evolutionary stages *per se*. With the help of the notion of lexical stickiness, we abstracted away from concrete word-word associations an innovative measure of how strongly words tend to occur in *n*-grams in general. At this level, we did not find a clear stage-to-patterning mapping either, but the results from the cluster analysis with two clusters emerging for the spoken corpus parts — Hong Kong English and British English versus Indian English and Singapore English — can be explained by plausible and powerful linguistic factors, especially different degrees of exonormative orientation and different degrees of influence from written norms. In the light of our findings and the interpretations that we offered, we would like to stress the importance of the level of descriptive granularity for the identifiability of mappings of evolutionary stages to linguistic patternings.

With regard to methodological issues of *n*-gram analyses, our findings make it very clear that one must exercise great care when differently homogeneous parts of one corpus, or different corpora, are compared. More importantly even, we need to be aware of the fact that our understanding of *n*-grams is still rather limited in spite of the increasingly large number of studies using them. This is apparent in several regards:

- We do not know which *n* is best for which purpose or whether it is in fact useful to decide on any one *n*, and we have proposed a way to address this issue.
- We do not know how to quantify the collocational strength of *n*-grams: most studies (and software) uncritically use statistics such as *MI* although these will typically be wrong. This is because they are based on the assumption of complete conditional independence, whereas it is quite clear that it does not make much sense to compute the collocational strength of *in spite of* on that assumption since the probability of *of* after the bigram *in spite* is very different from that of the product of the probabilities of *in* and *spite* separately. Again, we have attempted to outline an alternative approach.
- We do not know either whether the measure of lexical gravity that includes type frequencies is not in fact better suited as a collocational measure. Apart from the theoretical arguments mentioned in the present paper, there is also first empirical evidence. Gries (2010a) compares cluster analyses of the 19 sub-registers of the BNC Baby based on *G*-values of 2-grams to cluster analyses resulting from a more standard approach involving collocational *t*-scores. It turned out that the *G*-values replicated the structure of the registers and sub-registers of the corpus perfectly, whereas the *t*-score did not.

In spite of this *prima facie* evidence in favor of *G*, even *G* can probably be improved. On the one hand, the formula for *G* does not take the distribution of the type frequencies into consideration. To return to the example mentioned in Section 1.3, if $b=900$, i.e. there are 900 occurrences of *x* but not *y*, then all regular measures use the number 900 for the subsequent computation regardless of whether these 900 tokens consist of 900 different types or of two different types. The gravity measure, by contrast, takes this type frequency of two into consideration, but even *G* does not make a difference between two types with the token frequencies 450 and 450 and 2 types with the token frequencies 890 and 10, something that entropy can probably be used for. It should be obvious that even though ready-made software can now generate *n*-grams, there is a lot more that needs to be done to arrive at a better understanding of their distributional characteristics, and also their linguistic properties — an issue that has figured prominently in Biber's (and colleagues') most recent work on *n*-grams in academic writing (cf., e.g., Biber et al. 2004, Biber 2009, Csomay & Cortes 2010).

With regard to lexical stickiness, we are tempted to think that this is an interesting notion to explore. We can envisage applications not only in the general domain of collocation studies, but also more theoretically and/or didactically relevant applications. For example, if words are sticky, this by definition entails that their use will be governed more by the idiom principle than by the open-choice principle. Is it therefore possible to use this notion — in whatever exact way it will be operationalized — to quantify the position of words or lexical units on the proverbial cline from open-choice to idiomaticity? And can instruction in a foreign-language teaching context benefit from diagnostics that reveal how much words prefer to be used in collocations because these must then be taught? Or can some form of stickiness value indicate which words are particularly worthy of collocational study, much in the same way that Mason's (1999) approach of gravity can indicate which slots around a word are particularly variable?

Apart from these potential domains of application, future research should also delve more deeply into the question of whether words display different degrees of stickiness across varieties of English: given the fundamentally different contexts of language acquisition (and learning) in countries with English as a native language and English as an institutionalized second-language variety, it is not too far-fetched an assumption that the degree to which words tend to be used in larger (semi-)preconstructed units, i.e. their lexical stickiness, is highly variety-specific. In our view it makes sense to include the lexical stickiness of words in the set of lexicogrammatical variety markers to be analyzed in more detail in future corpus-based research into New Englishes.

Notes

[*] The order of authors is arbitrary. The present paper goes back to a paper presented at ICAME 30 at Lancaster University. We thank the participants as well as two IJCL reviewers for comments. We also thank Rosemary Bock for checking and proof-reading the manuscript. The usual disclaimers apply.

[1] See also Manning & Schütze (1999, in particular Chapters 5–6), Jurafsky & Martin (2008, in particular Chapter 4), and Crossley & Louwerse (2007) for many more examples.

[2] Following Biber (1995: 9f.), we use 'register' and 'genre' interchangeably and define register and text type as a situationally/communicatively-defined category and a linguistically-defined category respectively.

[3] It needs to be noted, however, that it would be more appropriate to compare postcolonial varieties of English with the diachronic variants of British English that were once transported to the new colonial territory (e.g. British English of the 17th to 19th century in the Indian context) rather than British English of the 1990's (cf. Hoffmann & Mukherjee 2007).

[4] In a similar vein, Olavarría de Ersson & Shaw (2003: 138) argue that "[v]erb complementation is an all-pervading structural feature of language and thus likely to be more significant in giving a variety its character than, for example, lexis".

[5] One of the reviewers was concerned that the sizes of the corpus parts used here may be too small for meaningful analysis. This is indeed a possibility but given that Gries (2006, 2009) and Gries et al. (under revision) worked with corpora and corpus parts of the same size and achieved robust results, we assume for now that the corpus (part) sizes are going to be sufficient.

[6] All text processing operations, statistical calculations and plots were done with R (R Development Core Team 2009).

[7] These and all cluster analyses below use non-centered Pearson as a measure of similarity of two vectors *x* and *y* and Ward's method as an amalgamation rule.

References

Arnon, I. & Snider, N. 2010. "More than words: Speakers represent multi-word sequences". *Journal of Memory and Language*, 62 (1), 67–82.

Baayen, R. H. 2001. *Word Frequency Distributions*. Dordrecht/Boston/London: Kluwer.

Bannard, C. & Matthews, D. 2008. "Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations". *Psychological Science*, 19 (3), 241–248.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. & Gildea, D. 2003. "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation". *The Journal of the Acoustical Society of America*, 113 (2), 1001–1024.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing". *International Journal of Corpus Linguistics*, 14 (3), 275–311.

Biber, D., Conrad, S. & Cortes, V. 2004. "If you look at …: Lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25 (3), 371–405.

Biber, D., Csomay, E., Jones, J. K. & Keck, C. 2004. "A corpus linguistic investigation of vocabulary-based discourse units in university registers". In U. Connor & T. A. Upton (Eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, 53–72.

Bolton, K. 2008. "English in Asia, Asian Englishes, and the issue of proficiency". *English Today*, 24 (2), 3–13.

Cavnar, W. B. & Trenkle, J. M. 1994. "*N*-gram-based text categorization". *Proceedings of SDAIR-94*, 161–175.

Crossley, S. A. & Louwerse, M. 2007. "Multi-dimensional register classification using bigrams". *International Journal of Corpus Linguistics*, 12 (4), 453–478.

Csomay, E. & Cortes, V. 2010. "Lexical bundle distribution in university classroom talk". In St. Th. Gries, S. Wulff & M. Davies (Eds.), *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 153–168.

Daudaravičius, V. & Marcinkevičienė, R. 2004. "Gravity counts for the boundaries of collocations". *International Journal of Corpus Linguistics*, 9 (2), 321–348.

Goldberg, A. E. 2006. *Constructions at Work: On the Nature of Generalization in Language*. Oxford: Oxford University Press.

Greenbaum, S. (Ed.) 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon.

Gries, St. Th. 2006. "Exploring variability within and between corpora: Some methodological considerations". *Corpora*, 1 (2), 109–151.

Gries, St. Th. 2010a. "Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora". Proceedings of *Corpus Linguistics 2009, University of Liverpool, 20–23 July 2009*. Available at: http://ucrel.lancs.ac.uk/publications/cl2009/404_FullPaper.doc (accessed July 2010).

Gries, St. Th. 2010b. "Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily …". *International Journal of Corpus Linguistics*, 15 (3), 327–343.

Gries, St. Th., Newman, J. & Shaoul, C. Under revision. "*N*-grams and the clustering of genres".

Hoffmann, S. & Mukherjee, J. 2007. "Ditransitive verbs in Indian English and British English: A corpus-linguistic study". *Arbeiten aus Anglistik und Amerikanistik*, 32 (1), 5–24.

Hopper, P. J. & Traugott, E. C. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Jurafsky, D. & Martin, J. H. 2008. *Speech and Language Processing*. 2nd ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

Kachru, B. B. 1983. *The Indianization of English*. New Delhi: Oxford University Press.

Kachru, B. B. 2005. *Asian Englishes: Beyond the Canon*. Hong Kong: Hong Kong University Press.

Kachru, Y. & Smith, L. E. 2008. *Cultures, Contexts, and World Englishes*. New York: Routledge.

Kilgarriff, A. 2001. "Comparing corpora". *International Journal of Corpus Linguistics*, 6 (1), 1–37.

Kita, K., Kato, Y., Omoto, T. & Yano, Y. 1994. "A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria". *Journal of Natural Language Processing*, 1 (1), 21–33.

Leech, G. & Fallon, R. 1992. "Computer corpora: What do they tell us about culture?". *ICAME Journal*, 16, 1–22.

Li, D. C. S. 1999. "The function and status of English in Hong Kong: A post-1997 update". *English World-Wide*, 20 (1), 67–110.

Li, D. C. S. 2009. "Towards 'biliteracy and trilingualism' in Hong Kong (SAR): Problems, dilemmas and stakeholders' views". *AILA Review*, 22 (1), 72–84.

Mair, C. 2007. "British English/American English grammar: Convergence in writing — divergence in speech?". *Anglia*, 125 (1), 84–100.

Manning, C. D. & Schütze, H. 1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Mason, O. 1999. "Parameters of collocation: The word in the centre of gravity". In J. Kirk (Ed.), *Corpora Galore: Analyses and Techniques in Describing English.* Amsterdam: Rodopi, 267–280.

McArthur, T. 2003. "English as an Asian language". *English Today*, 19 (2), 19–22.

McDonald, S. A. & Shillcock, R. C. 2003. "Eye-movements reveal the on-line computation of lexical probabilities during reading". *Psychological Science*, 14 (6), 648–652.

Memushaj, A. & Sobh, T. M. 2008. "Using grapheme *n*-grams in spelling correction and augmentative typing systems". *New Mathematics and Natural Computation*, 4 (1), 87–106.

Mesthrie, R. (Ed.) 2008. *Varieties of English, volume 4: Africa, South and Southeast Asia*. Berlin: Mouton de Gruyter.

Mesthrie, R. & Bhatt, R. M. 2008 *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Mukherjee, J. 2010. "Corpus linguistics versus corpus dogmatism — *pace* Wolfgang Teubert". *International Journal of Corpus Linguistics*, 15 (3), 370–378.

Mukherjee, J. & Gries, St. Th. 2009. "Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English". *English World-Wide*, 30 (1), 27–51.

Mukherjee, J. & Hoffmann, S. 2006. "Describing verb-complementational profiles of New Englishes: A pilot study of Indian English". *English World-Wide*, 27 (2), 147–173.

Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam/Philadelphia: John Benjamins.

Olavarría de Ersson, E. & Shaw, P. 2003. "Verb complementation patterns in Indian Standard English". *English World-Wide*, 24 (2), 137–161.

Orasan, C. & Krishnamurthy, R. 2002. "A corpus-based investigation of junk emails". *Proceedings of LREC 2002*, 1773–1780.

R Development Core Team. 2009. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna. Available at: http://www.R-project.org (accessed July 2010).

Reali, F. & Christiansen, M. 2007. "Processing of relative clauses is made easier by frequency of occurrence". *Journal of Memory and Language*, 57 (1), 1–23.

Schneider, E. W. 2003. "The dynamics of new Englishes: From identity construction to dialect birth". *Language*, 79 (2), 233–281.

Schneider, E. W. 2007. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

Simpson-Vlach, R. & Ellis, N. C. 2010. "An Academic Formulas List: New methods in phraseology research". *Applied Linguistics,* 31 (4), 487–512.

Solan, Z., Horn, D., Ruppin, E. & Edelman, S. 2005. "Unsupervised learning of natural languages". *Proceedings of the National Academy of Sciences*, 102, 11629–11634.

Stefanowitsch, A. & Gries, St. Th. 2003. "Collostructions: Investigating the interactions of words and constructions". *International Journal of Corpus Linguistics*, 8 (2), 209–243.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.

Trudgill, P. & Hannah, J. 2002. *International English: A Guide to the Varieties of Standard English*. 4th ed. London: Arnold.

Underwood, G., Schmitt, N. & Galpin, A. 2004. "The eyes have it: An eye-movement study into the processing of formulaic sequences". In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam/Philadelphia: John Benjamins, 153–172.

Wiechmann, D. 2008. "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias". *Corpus Linguistics and Linguistic Theory*, 4 (2), 253–290.

Xiao, R. 2009. "Multidimensional analysis and the study of world Englishes". *World Englishes*, 28 (4), 421–450.

Author's addresses

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106–3100
United States of America

stgries@linguistics.ucsb.edu

Joybrato Mukherjee
Department of English
Justus Liebig University, Giessen
Otto-Behaghel-Str. 10B, 35394 Giessen
Germany

Mukherjee@uni-giessen.de