The World English Model revised: Definite article use in ICE-GB and ICE-HK

Steve Disney
University College Plymouth Marjon

Abstract

This paper presents a cross-genre study of patterns of definite article uses in sections of ICE-GB and ICE H-K. The data analysis focuses on the timed student essay (TSE) components of each and finds differences in use types in ICE-HK/TSE than in ICE-GB/TSE based on the Quirk et al (1985) criteria. As the function of student essays is to display knowledge and, given that the conventions of different environments affect an individual's language choices, patterns of usage could be seen as a reflection of cultural demands. I conclude that there is a difference in knowledge display strategies between HK students' writing and British students' writing. The former prefers to overtly display knowledge, whereas the latter use a more indirect display strategy. I relate this research to Kachru's (1986, 1985: 12) concentric 'inner and outer circles' model and Yanu's (2001: 124) revised model, which accounts for register. I present a representation of the WE model that combines both models.

Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching, Vol. 4: Papers from LAEL PG 2009
Edited by Grace Bota, Helen Hargreaves, Lai Chia-Chun & Rong Rong

Introduction

English is the language of choice for many people whose first languages are not mutually intelligible, reflecting the dominance in world trade and power of first the British colonial system and now the Americans. Estimates of the number users in the world range from 340 million (Crystal 1995: 109) upwards, but this largely depends on how much English one needs to use in order to qualify. While most of the English of any given user in the world is intelligible to any other, there are many differences noted both between individual users and groups of users. As a result, there are countless descriptive and theoretical studies and projects on the use of English across different registers, genres, times and peoples from many different approaches. One of these is the ICE project, a collection of 1 million word corpora from different 'Englishes' using a balanced source set of naturally occurring data. The project was originally conceived by Sidney Greenbaum at University College London with the principle aim of providing 'the resources for comparative studies of the English used in countries where it is either a majority first language [...] or an official additional language' (Greenbaum 1996: 3). It is in this spirit that the present study is conducted, focussing on some differences between data in ICE-GB (Great Britain) and ICE-HK (Hong Kong). This paper begins by outlining some of the issues involved in the comparative study of the use of English in the world, with particular reference to Hong Kong. It goes on to describe a quantitative study of definite article usage in the selected sections of the corpora, and concludes with a suggested revision of previously proposed models of the nature of World Englishes.

English in the World

The spread of English and its status outside native speaker (NS) countries is the subject of much debate, i.e. whether it is a "variety", a "dialect" or "learner English". Widdowson, claiming to be provocative in order to widen debate, says English is a 'stabilized and standardized code leased out on a global scale, and controlled by the inventors, not entirely unlike the franchise for Pizza Hut' (1997: 140). The situation in 1980s post-colonial India led Kachru (1985: 12, 1986) to propose the concentric "inner and outer circles" model (fig. 1) of respectively old, new "contact" varieties, and a further circle outside these for EFL English.

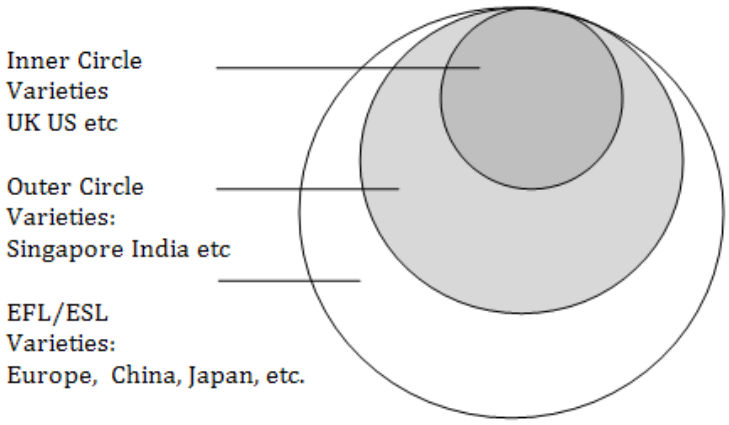

Fig. 1 Kachru's Circles of English

Because register is unaccounted for and the borders between the circles are more graduated than this, Kachru's model is somewhat limited, but the status debate itself remains valid, because the exported language adapts to local needs, influenced by local social and psychological variety (Widdowson, 1997: 137). Indeed, variation is reflected even in the individual, and speakers may feel the language they use is "theirs", e.g. EFL learners and NSs alike may apologise for "*my* bad English". It is unsurprising, then, that countries where English has official or semi-official status claim the English used there is not a dialect of some external "standard English" (SE), but is an independent "variety". Native variety status can be determined when it exhibits its own accent, history, literature, idiomaticity and localised, standard-setting grammars (Butler, 1997, cited in Bolton, 2000: 277). This issue is inevitably coloured by politics and, echoing the old "dialect debate", Halliday (2003: 406) reminds us that 'as linguists we have always insisted that a standard language was just another dialect, but one that happened to be wearing a fancy uniform'.

A mutually intelligible global language would seem to need to recognise some "standard" version, if only to avoid misunderstanding, but this is also politically charged. Yanu (2001: 129), promoting the Quirkian notion of a designed 'Nuclear English', says that if possible 'English for global use should be dissociated from the norm of any English-speaking society'. More usefully, Crystal (2001: 57) argues that a 'linguistically healthy world' will recognise both a standard variety and local varieties, and '[a] philosophy of diversity, recognizing the importance of hybridization, does not exclude the notion of a standard'. One unofficial standard he notes is "World Standard Printed English" (WSPE) with a spoken standard to evolve, perhaps internet prompted. Crystal (2001: 57) reports that he has reluctantly capitulated to using the term 'Englishes' because, 'if a community wishes its way of speaking to be a "language", and if they have the political power to support their decision, who would be able to stop them doing so?' Widdowson sums up the debate saying that,

> [a] dialect presupposes a language it is a dialect of. A code which declares independence is no longer a dialect but a language in its own right. People in Durham or Norfolk are not likely to declare independence. People in Ghana

> and Nigeria are. They may well wish to appropriate the language and make it their own. […] They are not dialects, they are something else. Something less continuous and dependent (Widdowson, 1997: 141-3).

Crucially, he says it is not 'the English language' that was exported to India, Hong Kong and Nigeria and spread through the world, but certain registers for certain purposes for certain people, those of power, commerce and science, rather like Norman French in Britain. These autonomous registers '…guarantee specialist communication within global expert communities. And this [...] is what most people are learning English *for*. It is not to indulge in social chat with native speakers' (Widdowson, 1997: 144). Yanu (2001: 124) revised Kachru's WE model to account for register, and to remove the value judgments inherent in the inner/outer circle distinction.

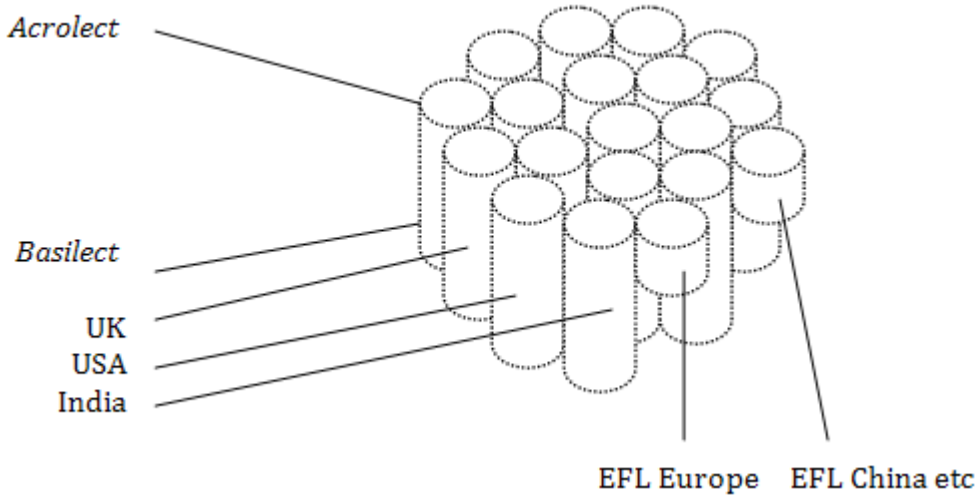

Fig 2 Yanu's (2001) Revised WE model

He used cylinders to denote each variety, with the acrolect at the top stretching to the basilect at the bottom to distinguish use with dotted lines to represent indistinct borders due to similarities. However, this model claims that EFL varieties do not extend into the basilect, which is disputable, as learners do indeed indulge in social chat, even if this is not the purpose of their study of the language. On the other hand, it could be claimed that the language they use in conversation is commonly more similar to that they use in the acrolect, thus making them sound more formal than they intend in conversation. I return to this issue briefly in the conclusion.

A new English variety arises when local needs result in expansion from acrolect registers into mesolect and perhaps basilect. This accounts for their core similarities and also for observed differences, i.e. acrolects exhibit fewer dissimilarities across varieties than basilects. In multi-lingual countries such as India and some African nations, English may be a social 'bridging language' other than just for trade, government and medicine and because the issues involve national identity it is difficult to accept anything other than Crystal's (2001) view, and accept the term 'Englishes'. Elsewhere, the spread through registers has differing effects,

from supplanting indigenous languages almost entirely (e.g. Australia), to producing a Creole (e.g. Jamaica). However, in Europe and Cantonese speaking Hong-Kong for example, people may be unlikely to use English outside the acrolect register domain (Bolt and Bolton, 1996: 201) and variety status is more problematic to establish. Here, an EFL/ESL distinction may be more appropriate, based on the language's sociological status and is 'a question of assigning [linguistic features] to particular sociocultural contexts of language use' (Benson, 2000: 379). This view clearly supports Widdowson's register analysis. It is the study of such variation that is a prime motivation of the ICE project.

This paper focuses on ICE-GB and ICE-HK. There is a key difference in the status of English in these countries; while in the UK, English is almost entirely an "inner circle" natively acquired language, in Hong Kong, English is an almost entirely taught and learnt variety (Bolt and Bolton, 1996: 199). However, British English is the target variety there and HK English in fact represents the most target-like English that Cantonese EFL learners in mainland China, are likely to achieve. Indeed, Yow (2001: 193 cited in Bolton, 2002) reports that it is an overt target of the Guangdong Education Commission, who wish to 'equip Guangdong students […] with the same command of English as their counterparts in Hong Kong and other Southeast Asia countries'. It is therefore essential to study the language produced by their HK student counterparts, insights from which may be of use to others.

World Englishes (2000/3) was entirely devoted to 'English in Hong Kong', and although the discussion on its status was inconclusive, 'Hong Kong English' was widely referred to as though it were an established language variety. Due to Hong Kong's colonial status, English has been an official language of Government, Education and law since 1841, and is dominant in business. However, Bolt and Bolton, writing even before the political handover to China, report that 97% of the HK population are Cantonese speakers (1996: 197), and fewer than 7% self report using English 'well or very well' (1996: 200). In reality, while remaining a language of instruction, government and law, it is no longer *the* language used thus and it is Chinese Government policy to increase use of Chinese, thus reclaiming the acrolect discourse domain. Bolt and Bolton feel that English is going to struggle to remain a medium of instruction, while English broadcasters are being supplanted by Chinese and foreign channels, and circulation of English language newspapers is falling rapidly (Yeun-Ying' 2000).

The International Corpus of English (ICE) project

Corpus Linguistics as a research methodology aims to provide examples of real usage events as the basis for demonstrating the variety and patterning in any language. Thus it is 'usage-based' in Langacker's (1987 & 2008) terminology. Corpora have been collected that includes large collections, primarily used for lexical research, such as the COBUILD and BNC projects, and smaller, one million word,

grammatically parsed corpora such as ICE[1] that this study uses. Each ICE corpus consists of samples of educated native speaker (NS) data, with multi-genre samples of conversation, monologues, and written sources including academic, letters and fiction, across many of the global varieties. There is a subproject, ICE-ICLE, collecting data from EFL/ESL environments, but the Hong Kong[2] data (ICE-HK) is not included under ICE-ICLE, despite the debate on the status of English there.

For ICE-HK, samples were collected from native Cantonese speakers who fulfilled criteria such as not having been educated in NS countries, although some were not HK natives. 95% of contributors fitted the profile, but because public English is dominated by NSs, there were problems reaching targets and unless there is a NS of English present, two HK Cantonese speakers will speak Cantonese. Bolt and Bolton (1996) feel their work may coincide with a 'high-water mark' in English use in Hong Kong. They also report difficulties tagging and parsing texts, as up to 40% of the data was 'non-standard', including code-switching, Cantonese interjections and idiom (Bolton, 1996: 211). These difficulties continue, so this study is based on an unparsed text version, and regrettably little analysis can be made at the phrase level on the corpus as a whole.[3]

Background

This study examines the usage of *the* in ICE-GB and ICE-HK, with an emphasis on how a speaker signals to a hearer the location of referents in their discourse domain. In the ICE-GB corpus, *the* is the most common word, and its "indefinite" counterpart *a/an* (henceforth *a*) is among the top ten. They are frequent in English largely because single common countable NP heads are extremely frequent and Modern English has a "rule" that 'a singular count noun can not be used without a determiner' (Hudson, 1992: 219).

At a purely descriptive level, one can broadly say that in the absence of another determiner, an article acts as a default, depending on certain pragmatically based selectional criteria. More precisely, there is an argument that *a(n)* is merely a filler, used when *the* is disallowed for some reason, i.e. it signals only a lack of what ever it is that *the* is used to signal. This is why this study considers only constructions with *the* and not *a*. There is much discussion and disagreement about the grammatical status of articles in terms of category membership, (e.g., Biber et al, 1999; Huddleston, 1984; Quirk *et al*., 1985). For some, they are adjectives (Gleason, 1965), for other pronouns (Jespersen 1933), for others they are determiners (Abney, 1987) and still others claim they are separate category (c.f. Spinillo 2005 who classifies them

[1] ICE-GB is hosted at the Survey of English Usage in the English Department at University College London, UK

[2] The ICE-HK project was funded in part by the Hong Kong Research Grants Council (RGC), Grant no. HKU 7174/00H

[3] Concapp and Wordsmith Tools were used for handling text data, and ICECUP was used for the parsed ICE-GB corpus.

with *every*).[4] These are based on the assumption that word classes are definable in terms of their properties and syntactic behaviour. However, from a modern Cognitive Linguistic and (Radical) Construction Grammar perspective such a debate is quite meaningless, due mostly to the fact that syntactic relations are denied (Croft, 2001). Croft shows how in fact it is impossible to categorise language specific word classes in this or any other way. In line with this approach, the key to understanding Article usage, like all other so-called 'function' words in English lies in the constructions and Usage Events in which they are used, not in the notion of 'article' as some sort of discrete or even universal grammatical category. The approach would in addition argue that there is no 'grammar rule' with respect to the use of the article and that in a particular usage event what an article signals is grounded in that particular shared discourse space (see also Goldberg, 1995; Croft, 2001; Langacker, 2008).

The debate remains, however, that articles are used to signal *something* between speaker and hearer, and use is highly conventionalised. In quite loose, but established, category terms *a* contrasts with both the English plural *-s*, i.e. [± singular] and with *the*, e.g. [± indefinite]. This has the effect of restricting its occurrence to cases that are both +singular and +indefinite. Both of these concepts are shown to be cross-linguistically salient categories and form distinct areas of universal conceptual space (Croft, 2002; Haspelmath, 2003; Croft and Poole 2004)

The distinction in English between *a* and *the* cannot, however, be expressed in any simple fashion, although there are many attempts. The main theorists also make this point, e.g. Lyons wishes to classify *a* and *the* together for intuitive reasons, but says that the real distinction is 'difficult to pin down' (1999: 106). Some descriptive analyses are based on 'definiteness', 'specificity' or 'uniqueness', or combinations thereof, while other approaches relate syntactic patterns to semantic distinctions within a specific theoretical framework. Halliday and Hasan describe *the* in its anaphoric cohesive role, and state 'the definite article has no content' (1976: 71). Christopherson (1939: 71) says 'It is found that *the* has the marking of *familiarity*, while *a* is a mark of *unity'*. This definition is more debateable considering the example *Beware of the dog*, which has *the* used with an unfamiliar dog. Some notion of shared discourse space seems essential, an argument that adds further weight to the view that 'grammaticality' is not contained entirely within the syntax of a language (Hawkins, 1978: 91), a point that is fundamental to the Construction Grammar

[4] However, the system is complicated by a posited "zero" article Ø the main role of which appears to be to lend weight to the obligatory determiner hypothesis. Much disagreement and inconsistency exists on this in the literature with some commentators saying "the zero article is used" (Quirk *et al*., 1985: 274), while others refer to "bare" plurals (Lyons, 1999: 190). Biber et al (1999: 260) report that "it is customary to recognise a zero article", but add (1999: 261) "arguably some of these cases should be analysed as involving neutralisation of article distinctions, rather than cases of zero article". The "zero article" is excluded from this study as it may be better considered an absence of an article, rather than a presence of a nothing. Also, it is difficult to search a corpus for that which is not there, and equally difficult to argue that once found, such a nothing has semantic content or pragmatic function. Lack of *a* or *the* referred to with Ø below does not imply "existence of a zero article".

approaches cited above and indeed to the approach of Hawkins (1978) who is the most cited work on the topic of definiteness in English, and the starting point for many other commentators, e.g. Lyons (1999). In arguing against semantic approaches based purely on the concept of definiteness and indefiniteness, Hawkins (1978: 89) says, articles 'cannot be explained or even discovered in abstraction from pragmatics'. In answer to claims that they give a value based on relative 'uniqueness', he says this is '… just a single manifestation of a more general regularity: inclusiveness within pragmatically defined parameters' (Hawkins, 1978: 17). The reason 'inclusiveness' had been missed and 'uniqueness' relied on is, he claims, the mistaken assumption 'that uniqueness is an absolute rather than a relative notion […] independent of any pragmatic considerations' (Hawkins, 1978: 161) i.e. a given referent may not necessarily be unique in the real world, but still be unique within interlocutors' shared knowledge and/or discourse space.[5]

The fact that there are many conventionalised uses is inevitable given their ubiquity and longevity. Hawkins concludes that 'uniqueness' is also not a part of the meaning of *the*, but 'results from a fusion of the meaning of the definite article with singularity or oneness' (Hawkins, 1978: 158). For him, the articles are pragmatically motivated. He says they provide an instruction to the hearer to locate an entity in the shared discourse or the universe, or to introduce such an entity to the discourse, a clear pragmatic particle. Hawkins argues that even truth / false logic may fail because an assertion cannot be judged either true or false if it fails at another level. For example, the assertion *A Prime Minister of England is bald* can be neither true nor false as it is *pragmatically* unacceptable i.e. it does not exclude the possibility of there being more than one Prime Minister. Hawkins (1978: 89) concludes that

> …logical meanings cannot be successfully discovered without constant reference to the full range of usage possibilities. One cannot ask the NS to corroborate the existence and uniqueness claims made by definite descriptions. These are abstract and problematic notions in themselves. What does it mean for something to exist? What are the parameters relative to which objects are unique? (Hawkins, 1978: 89)

Hawkins offers the categories for definiteness summarised in table 1 and which are also used by Lyons.

Label	**No.**	**Description**	**Example**
Generic Use		All generic reference, inc gerunds	
Anaphoric	1	Second mention etc	a bucket: the bucket a lathe: the machine
	2	a) Visible Situation	Pass me the bucket

[5] Hawkins further claims that one reason the articles have no core semantics (1978: 13) is because neither appeared in the Germanic source languages, *a* being an enclitic of numeral *one*, and *the* being derived from the demonstratives. In line with standard grammaticalisation theory (e.g. Bybee, 2003) the result of these (independent) developments are more abstract and more schematic than their source forms.

Situational:		b) Immediate Situation	Beware of the Dog.
	3	**Larger Situation:** a) Shared specific knowledge	I'm going to the match.
		b) Shared general knowledge: The situation is the trigger.	At a match: Who is the referee?
Associative Anaphoric	4	Pragmatic Set of possibilities: Previous NP is trigger	a: I'm going to the big game. b: Ah, who is the referee?
Unfamiliar with explanatory modifiers	5	The 2nd part defines the definite ref. a) Establishing Relative Clauses b) Associative clauses c) NP Complements d) Nominal Modifiers	The woman who he met was rude. The start of the game was delayed. London is buzzing with the rumour that the PM lied. I don't like the colour red.
Unexplanatory	6	Logical use.	The same train as always. The first man on the moon.

Table 1 Hawkins' (1978) model

The philosophical tenor of the Hawkins model is mirrored by Quirk *et al.* (1985), Biber *et al.* (1999) and Huddleston and Pullum (2001). For example, the last talk of 'existential presuppositions' (1985: 369), while Quirk *et al.* describe a definite/indefinite paradigm. They also quite neatly manage to include 'uniqueness', and 'shared knowledge', in the highly quotable

> The definite article *the* is used to mark the phrase it introduces as definite, i.e. as "referring to something which can be identified uniquely in the contextual or general knowledge shared by speaker and hearer (Quirk *et al*., 1985: 265).

However, the large number of footnotes reflects the complexity of the discussion, e.g.

> In practice, since a speaker cannot always be sure of the hearer's state of knowledge, use of *the* involves a certain amount of guesswork. In fact, in some cases the assumption of shared knowledge is a palpable fiction (Quirk *et al*., 1985: 266).

While the Hawkins model is very complete, the conflated framework found in Quirk et al is very similar and is preferred herein. Although because it is just a descriptive list of uses, not a theory of use, it is somewhat lacking in explanatory power. It is, on the other hand, better able to cope with the sort of usage categorisation I present below. Table 2 summarises the Quirk et al framework. The equivalent categories of the Hawkins analysis are added in the final column for comparison.

a. immediate situation	*The roses* are very beautiful. (said in a garden)	**2**
b. general knowledge	the Prime Minister, the sun	**3/ii**
c. direct anaphoric ref.	John bought a TV and a video recorder, but he returned *the video recorder*.	**1**
d. indirect anaphoric ref.	John bought a bicycle, but when he rode it one of *the wheels* came off.	**4**
e. cataphoric ref.	*The girls* sitting over there are my cousins.	**5**
f. sporadic reference	My sister goes to *the theatre* every month. What's in *the paper* today?	**3/i**
g. "logical" use	This is *the only copy*. When is *the next bus?*	**6**
h. generic meaning	No-one knows precisely when *the wheel* was invented.	**gen**

Table 2 Definite Article uses distinguished by Quirk et al (1985: 265, 282–7)

The definite article in ICE-GB and ICE-HK

This section presents a quantified analysis of use of the article *the* within the timed student essays sections of the ICE-GB and ICE-HK corpora. The analysis below uses the Quirk criteria, but retains the Hawkins 1-6 numbering for ease of cross- reference.

In brief, across the whole ICE-GB corpus *the* accounts for 5.48% of all words, a repeat rate (rr) of 1 in 18.25 words, while in the ICE-HK corpus it is 5.08% (rr=19.68). This is basically in line with Sand (2004: 290), who writes 'a quantifiable "underuse" of the definite article for varieties whose substrate does not contain definite articles cannot be substantiated'. He claims that there is a 'varied and text-type specific distribution of definite articles' (Sand, 2004: 290). The overall use difference found here is not great, but it is "quantifiable".

I focus here on uses of *the* in one comparable subcorpora, timed student essays (TSE). This particular sub-corpus has been chosen because, as Biber et al (1999:160) say, cognitive and time pressures in exams mean there is limited opportunity for writers to revise and re-write, i.e. student essays have a more 'on-line' characterisation. One result of this is that it may be a closer representation of users' linguistic knowledge than language produced, and subsequently extensively analysed and revised by the writer, in other written mediums.[6] This selection is of practical interest too, because if differences were found between the two sources, writers of such essays from Hong Kong or other Cantonese or even other Chinese speaking regions may find the results useful, should they wish their writing to more closely match the standard British English expected of the writers of the ICE-GB data.

The data in each of the TSE sources were manually classified according to type of use in line with table 2 above based on Quirk et al (1985). Some error rate in

[6] Conversation data could of course fulfil the same role, but presents its own problems in analysis, especially in the ICE-HK corpus where it is disjointed and exhibits frequent use of Cantonese. It is therefore not entirely clear how comparable the data sources are and at least with student essays, one may be sure of a controlled and comparable environment.

classification must be acknowledged, although this is low enough to have had a negligible effect on the results. There was an emphasis on consistency in the classifications, but some problems did arise, mainly connected with topic and unclassifiable instances. Firstly, the fact that in ICE-HK/TSE the word *starch* is the 16th most common word shows that the data sourcing was not as varied as would be hoped and closer topic matching between the sources would have been helpful. Secondly, some examples of poetry were supplied with an examination paper, encouraging the writer to use quotations in their essay, e.g. (1)[7], and some answers required examples, e.g. (2).

(1) … the words "jewel of *the* just" promote the then Christian ideal that only those who have led good lives will attain their place in glory… GB 18:136:3

(2) a) Error in this stage result from mis-selection of affix info, for example: HK 11:132:

b) "I put *the* steaks into *the* freezer" HK 11:133:1

Other topics required titles of places with *the*, e.g. *The USA*, books, music etc. and titles of theories etc. Cases such as these were removed from the examined data as they are not instances of the writers' own use.

The ICE-HK/TSE data also presented some unique problems where examples were either not readily comprehensible or were clear cases of error e.g. (3).

(3) *The* improve *the* gel strength modification of starch chain can be use. HK 19:36:1

In all, 303 instances were discounted from the ICE-HK/TSE data and 101 from ICE-GB/TSE and after removal table 3 shows that there is a 13% higher incidence rate of *the* in ICE-HK/TSE than in ICE-GB/TSE, accounting for nearly 1% more of the total number of words in the sub-corpus.

	Total words	Total THE	repeat rate	% of all words
ICE-GB/TSE	21262	1367	**15.55**	**6.43%**
ICE-HK/TSE	24498	1809	**13.54**	**7.38%**

Table 3 Revised figures for the use in TSE data

It has not been possible to analyse NP rates in ICE-HK/TSE, although a cursory examination suggests NPs there tend to be shorter and more frequent than in ICE-GB/TSE. If this were true, a higher overall use of articles would be expected because articles are a function of NPs. However, because in the ICE written data one text unit

[7] All ICE references are taken from W1A sources and are here prefixed either GB for ICE-GB references and HK for ICE-HK references. In the interests of clarity and parsimony, some irrelevant parts of some source text units are omitted and replaced by "…" . The instance of *the*, or a phrase, under consideration in each case is in added italics.

basically corresponds to one sentence, the number of text units in the two sources is comparable. Table 4 shows that ICE-HK/TSE has 82 *the* tokens per 100 text units while ICE-GB/TSE has 83. Because rates are nearly identical per unit, no overall difference can be claimed, in contrast to Sand (2004: 290) above.[8]

	Text units	words per unit	*#the per 100 units*
ICE-GB/TSE	1136	18.7	**83**
ICE-HK/TSE	1497	16.3	**82**

Table 4. *The* use by text unit

The overall distribution by type of use is shown in table 5, and reveals some clear differences between sources.

Type of *the*	**ICE-GB**	**% of *the***	**ICE-HK**	**% of *the***
Generic	72	**5.27%**	169	**9.34%**
1. Anaphoric	217	**15.87%**	424	**23.44%**
2. Immediate Sit.	67	**4.90%**	63	**3.48%**
3. Gen. Knowledge	201	**14.70%**	96	**5.31%**
4. Associative	184	**13.46%**	333	**18.41%**
5. Cataphoric	514	**37.60%**	586	**32.39%**
6. Logical Use	112	**8.19%**	138	**7.63%**

Table 5 Percentage use of the by type in timed student essays

Overall, type 5, cataphoric reference, (i.e. ***the*** *x of y*) is the most used, at about a third of all use in both sources while type 2, immediate situation, is the least used in both sources. This is in line with Biber (1999: 266).[9]
However, beyond this, usage by type varies quite starkly, with each source showing marked preferences. To anticipate the explication below, the ICE-HK/TSE data shows a preference for anaphoric reference, while ICE-GB/TSE appeals more to general knowledge. The analysis below focuses mainly on this particular finding.

Anaphoric reference

Taken together, direct and indirect anaphoric reference account for 42%[10] of all *the* use in ICE-HK/TSE and 29.5% in ICE-GB/TSE a . This means rates are 30% higher in HK/TSE, which is a significant difference (z-score = 7.43 $p<0.01$) and one which suggests a fundamental difference in the approach to the presentation of information between the sources. It seems that in the ICE-HK/TSE sample, entities are expressly

9 However, one claim that can be made is that the ICE-HK/TSE samples are clearly shorter than the ICE-GB/TSE, perhaps suggesting they are less complex.

[9] It should be noted that Biber et al (1999) classified up to 5% of their data as "uncertain", an option that is understandable with the amount of data they examined. For a small study like the current one, this was not really a problem but the effects that borderline cases may have on the results is noted as it arises.

[10] Numbers are rounded up to the nearest 0.5 for clarity.

introduced and subsequently referred back to anaphorically to a far greater degree than in the ICE-GB/TSE sample.

Type 1 Direct anaphoric reference

In the rates for Type 1 there is 33% more use in ICE-HK/TSE at 23.5% of all *the* use than in ICE-GB/TSE at 16% (z score = 5.4 $p<0.01$). The higher rates of lexical anaphora in the HK sample may be a further reflection of a high NP rate; the more NPs there are, the more often a given NP is likely to have to be referred to. Examples (4-5) show prototypical anaphoric uses, with the NPs co-indexed by lower case 'i'.

(4) Moreover if *a paw*$_i$ was tapped just before it was due to be lifted *the paw*$_i$ would then be reflexively lifted an increased amount… GB 16:60:2

(5) Syntactically speaking, the nature of say, *a verb*, will allow us to draw inference about what will be coming in that single sentence. HK 20:87:1
For instance, whether *the verb* is a transitive, intransitive or a ditransitive one will help us infer what comes next. HK 20:88:1

The difference in rates may be partly accounted for by the high rate of lexical repetition evident in the ICE-HK/TSE texts, where in an equivalent situation, an ICE-GB/TSE writer may prefer to refer by pronoun. A good set of parallel examples of this are provided in (6a-b) and (7a-b) below, but lexical repetition cannot account for all the difference.

(6) a) *The starch* can be converted into sweetener . HK 19:67:1
b) *The starch* remain part of corn plant can be used for alcohol production while the residue can be use as animal feed. HK 19:68:1

(7) a) *Folding* occurs in the upper crust (lithosphere) as the rock structure needs to be preserved. GB 20:114:5

b) However *it* may extend into the lower crust as in the case of deep seated isoclinal folds etc. GB020:115:5

Type 4 Associative anaphoric reference

This type of use is described by Hawkins (1978: 123) as being 'the most frequent use that is made of *the*' and as 'the most fascinating'. These data show that for TSE such use is not the most common with rates of 13.5% in ICE-GB/TSE and 18.5 in ICE-HK/TSE as a %? of total use of *the* (z score = 3.88 $p<0.01$). It remains to be seen if type 4 rates have a linear relationship with type 1 rates, but the ratio of use is very similar at approximately 5:4, (79% in ICE-HK/TSE and 81% in ICE-GB/TSE), with type 1 the more common in both.

As a central example (8) shows *the display* as indirectly anaphoric to the situation *museum*. Example (9) is particularly complex, where the reference to *the stage* is indirectly anaphoric to the situation *musical drama* and *the violin* to *the violinist*, which is in turn also indirectly anaphoric to the situation *musical drama*.

(8) At the Museum of Mankind I find the repetition of description as tiring and as the fabricated coldness of *the Arctic display*. GB 12:42:1

(9) … when *the violinist* finishes his playing the king grab *the violin* and smashed into pieces and as a result he became mad and howling off *the stage*. HK 14:58:1

In (10) the reference to *the males* is indirectly anaphoric to the mention of *a group of people in society* earlier in the text unit.

(10) But I wonder if this is the true reality or that such a family ideology is a distortion of the true picture, aiming at the rested interests of a group of people in society, in this context, *the males* HK 12:19:1

There were some problematic cases connected with the classification of type 4 uses, because there is a cline with type 3, larger situation/general knowledge use. This is discussed below, but it should be noted that as consistency of analysis was considered the prime concern, where a specific antecedent was not readily identifiable, a given instance was classified as type 3.

Type 2 Immediate situation

Use of this type is low in both sources at 5% in ICE-GB/TSE and 3.5% in ICE-HK/TSE (z score = 1.97 $p < 0.05$), but analysis did reveal some differences. Firstly, there are vastly more text internal references using *the* + N within the ICE-HK/TSE texts (#37) than the ICE-GB/TSE (#6):

(11) *The above examples* show clearly that the failure of agrarian reform is certainly a major factor in explaining environmental degradation… GB 13: 70:2

(12) From *the example*, we can see that speech error can help linguists to conjecture speech production model. HK 11:68:1

Secondly, the overall rates may be very similar, but most type 2 references in ICE-GB/TSE were produced by one writer, when writing about a poem that was supplied with the question paper. The first direct reference to the poem was in (13), and the next was as shown in (14), the following text unit.

(13) I think that this poem justifies his point. GB 18: 119:3

(14) The poem essentially about the poet's longing to escape his present existence. GB 18: 120:3

Both of these are considered to be pointing directly to the supplied poem, and not that (14) is anaphoric to (13). If it were anaphoric, one *particular* previous instance would need to be identified as the antecedent, which proved not possible, as they are merely co-referential to all other previous mentions of, e.g. *the poem*.

Type 3 larger situation / general knowledge use

The rate difference between sources is 15% ICE-GB/TSE and 5.5% in ICE-HK/TSE, so type 3 use is almost three times higher in ICE-GB/TSE (z score = 8.64 $p<0.01$). Central examples are given in (15) and (16).

(15) For example, when a doctor comes across a suspected child abuse case, he may need to inform *the* police or *the* social welfare department (…)
HK 12:51:1

(16) Such colonisation programmes are carried out in Amazonia but pose severe threats to *the* environment. GB 13: 68:2

Type 3: Problematic Cases

Most type 3 uses of *the* were unproblematic, being prototypical like these, but it was classifying this type that gave the most difficulty overall. For classification purposes, if no specific antecedent was readily identifiable within the student's actual answer paper to act as a trigger, the reference was classed as type 3, general knowledge.

The main problem is that in student essays the audience for the writer is a specialist in the subject being discussed and it is impossible for the analyser, a linguist, to know what information is or is not +hearer known [+HK] for the expert reader in, for example, the manufacture of bread (ICE-HK/TSE) or rock formations (ICE-GB/TSE), two major topics found in the data. The classification issue concerns whether to count another NP as an antecedent and classify the case in question as type 4, associative anaphoric, or classify the reference as type 3, general knowledge.

The problem arises in varying degrees of uncertainty, e.g. in (17) a possible antecedent for *the corn plant* is *corn grain* mentioned in this and the previous text unit, although one can assume that everyone knows of the existence of corn plants.

(17) Corn grain is very important because all *the* corn plant can be use without waste. HK 19:65:1

In contrast, in (18) the writer assumes that the reader already knows what *the classical form* consists of, but this knowledge may not be a part of the average person's general knowledge.

(18) The form of the 20th centuries British opera is mainly the form of *the* classical form. HK 14:111:1

Similarly, in (19) the writer makes an assumption that gastropods exist in the rocks being discussed and that they can be analysed in a certain way for a particular purpose, information missing from this linguist's real world knowledge, but not, the writer assumes, the examiner's. These types of inference are typical of type 3 uses.

(19) Similarly *the* gastropods can be used to help dating and correlation in the lower Paleozoic. GB 20: 95:3

Further examples show how complex some presuppositions are, and that interlocutors build a 'shared discourse universe' based on assumptions of [+HK]. In (20) there is a presupposition of the existence of 'victimhood', as an integral part of the concept of inequality. Such use relies on a presupposition of the reader holding [+HK] of a whole political philosophy, not merely a particular entity[11].

(20) This kind of 'false consciousness' has indeed generated gender inequality in which women are *the* victims. HK 12:37:1

Similarly, in (21) a whole body of economic theory is assumed [+HK] in the reader, in order for the writer to blithely refer to a *recession* that is both *current* and *worldwide*.

(21) Finally, *the* current world recession makes rapid progress for LDC's even more unlikely. GB 13: 34:1

There is then a cline between what is anaphoric and what is general knowledge with some borderline cases. Borderline cases are problematic for the statistical analysis and may weaken any claims. However here, because this categorisation issue occurred to a greater degree in the ICE-HK/TSE data, if in fact some instances currently classified as analysed here as type 3 were analysed as type 4 instead, the results would in fact *increase* the trend differences between sources, i.e. in ICE-HK/TSE there would be even more type 4 and less type 3 uses. As it is, the differences are still significant, as reported above.

Another issue concerning categorising between types 1 to 4, and possibly unique to student exam or term papers[12], is co-reference in an answer to a specific entity mentioned in the question. This is even more of a problem here because the exam questions and material are not supplied in the corpus. For example, if *the poet* in (22) co-refers to an entity in the question paper, e.g. if the paper asks 'How does the poet feel about *x* ?', it could be argued that the reference is type 2.

(22) It helps to clarify *the poet*'s ambiguous comments beforehand by giving an actual example of what he means. GB 18: 33:1

On the other hand, *the poet* in (22) could be anaphoric (type 1) to the question, or general knowledge (type 3) to the topic of *poetry* and merely co-referential to the question. It could also be indirectly anaphoric (type 4) to a mention of *the poem*. Therein lies the true problem facing the analyst: a lack of the compete context. Here, in such cases, the writer's first mention of *the poet* is classified as type 4, as it is more

[11] On the other hand, this may be a case of error by insertion and may have been meant to have a reading like "in which women are victims". It is not possible to ask the writer for clarification so the usage as it appears has been analysed and no assumption of error is made in ambiguous cases.

[12] Although Newspaper/magazine headlines may also cause this problem in an analysis.

consistent to argue [+HK] exists within the shared knowledge of a discourse universe (such as 'poetry') as introduced by the question, and is anaphoric to this. This contrast with the idea in type 3 where the writer relies on some notion of shared general knowledge that has no antecedent in the current discourse. Because this type of reference occurs more in ICE-GB/TSE than the ICE-HK/TSE, again, were they in fact to be classified as type 3 and not type 4, this would also increase the usage differences between the sources.

Type 5 cataphoric reference

The final major usage type is cataphoric reference, (23-24) where a noun's reference is completed by some form of post-modification. This is in fact the most common type of *the* use, accounting for 38% of *the* use in ICE-GB/TSE and 33.5% in ICE-HK/TSE (z score = 3.11 $p < 0.02$).

(23) Thus *the duration* of post traumatic amnesia is related to *the severity* of the injury. GB 16: 73:3

(24) *The output* of each stage becomes *the input* of the following one. HK 20:110:1

The 5% difference in type 5 usage between sources is accounted for almost entirely by the lower use in ICE-HK/TSE of the string *of + the*. This lends some support to the point raised above that NP structure in the HK data may be less complex than that of the GB samples.

Erroneous use of the in ICE-HK/TSE

One would expect there to be many errors in what is essentially an ESL variety, if not an EFL variety. There were two basic error types in ICE-HK/TSE, neither of which occurred in ICE-GB/TSE. These were errors by insertion, and errors by omission, although clearly this relies exclusively on researcher judgement. The error rate was 9.1% so there is a potential for misjudgements to have an effect on the overall results reported here. However, the erroneous cases were for mostly quite clearly a misuse. The most common was insertion of *the* but this is of no further interest in this paper.

Discussion

There are some other context contingent conclusions that are given above, but reservations are acknowledged about corpus design, data volume and the fact that although prototypes are abundant, borders between classifications are fuzzy. However, the most significant differences found were between the rates for both types of anaphoric reference, where combined use was 30% higher in HK/TSE ($p<0.01$), and reference to general knowledge [+HK], which is three times higher in ICE-GB/TSE ($p<0.01$). Clearly, the validity of any explanations offered relies on the

accuracy of the analysis and on the choices of exactly what comparisons to make. To this end, other referential similarities should not be ignored. There are two points to be raised here.

First is the conflation of [+HK] references (uses 3+4) as they are distinguished from other uses by being reliant on the assumption by the speaker of knowledge held by the hearer, even though the former is anaphoric and the latter is not considered to be so. When uses 3 and 4 are combined, ICE-GB/TSE 28.5% of the use and ICE-HK/TSE for 24%, suggesting ICE-GB/TSE writers rely more on assumptions of [+HK], but perhaps not to the extent that the figures for situational use alone would suggest and in fact gives a significance of only $p<0.05$.

Second is a conflation on the basis of [± deictic] (i.e. uses 1, 2 and 4) shown in figure 1. The rates for [+ deictic] are in ICE-HK/TSE at 45.5% and ICE-GB/TSE at 34.5% suggesting ICE-HK/TSE writers use more direct reference and less inference.

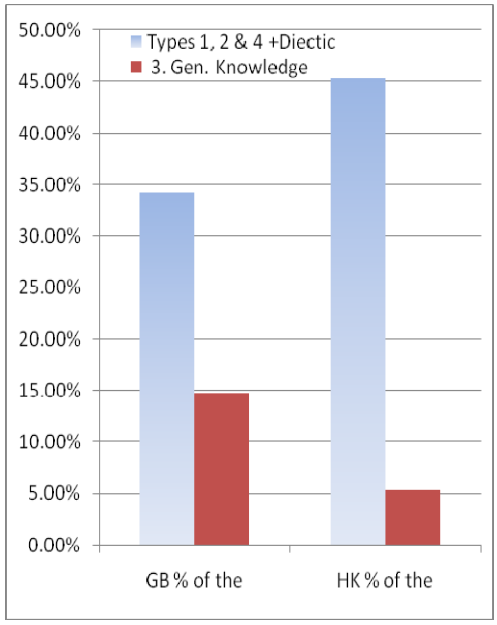

Figure 1. Conflated uses 1,2 &4 compared to use 3

So GB writers rely significantly more on assumptions of hearer knowledge and HK writers more on the explicit introduction of referents and following anaphoric reference to them. Because articles can be argued to be pragmatic particles, then an explanation must be found at the relationship level, i.e. the relationship between student writer and their interlocutors, the marker/lecturer. Clearly, one issue that affects linguistic features is the reason for producing a text. The purpose of student essays is very different from that of other genres, being primarily to display knowledge, rather than impart information. Further, the requirements and conventions of the particular institution will affect the language choices of the individual being assessed and any apparent patterns of usage could be seen as a reflection of such demands. There are different ways in which such knowledge display may be achieved, e.g. explicitly or assumed.

There are clearly many linguistic features in which such a pattern may be reflected e.g. Huebner's [-SR +HK] 'equivalence' use. Example (25) from ICE-HK/TSE shows a common way of explicitly stating knowledge for display purposes.

(25) The Conceptualize is divided into 2 group. HK 11:78:1
The first is message generation. HK-011:79:1
It plans the message. HK 11:80:1

This pattern affects type frequencies of *the* in the following way. If a reference is overt, a singular first mention of a noun may be with an indefinite article. Subsequent mentions with *the* will clearly count as 'direct anaphoric'. However, if knowledge of a referent is *assumed* by the writer, the first mention may be a type 3 use of *the* and there may be no need for a subsequent mention, thus lowering anaphora rates. Half way between these types is type 4 use, which is partly anaphoric and partly assumed knowledge.

This would suggest as a tentative explanation that there is what appears to be a systematic culture-specific knowledge display strategy in operation. That is, there is a preference among HK students to overtly display knowledge, whereas the British use a more indirect display strategy. It is not a claim of this paper that the British writers did not use such direct display strategies, merely that the Hong Kong writers use them significantly more and *vice versa* with indirect display strategies. Regrettably, a more detailed examination is not possible here, but would make a useful future research project and possibly be of great use to the growing numbers of students studying in UK universities from China and other Pacific Rim countries who also have no articles and may well pattern similarly.

Conclusion

This paper has described and compared use of *the* in ICE-GB/TSE and ICE-HK/TSE according to the Quirk et al (1985) and Hawkins (1978) paradigms, and has found many similarities in distribution and some significant differences. There are also, I believe, implications for the World English models outlined in section 2 where I expressed reservations of both models, and crucially that the revised Yanu model excludes conversation in EFL environments. It also not does not show the range of differences and similarities possible in varieties, e.g. at the acrolect level a proficient EFL learner may be close to WSPE, yet still very far from SE in a typical basilect register such as conversation or social letter writing. The fact that there are differences is nothing new, and indeed was the main motivation for the inception of the ICE project. As Kachru and Nelson (2001: 12) say 'that there are differences does not automatically imply that someone is wrong'. However, in representing the state of things graphically, one needs to prioritise and choose between e.g. usefulness and ideology. The strength of the Kachru model is the clear notion of central and non-central types relating to standards, while Yanu's model is more socially 'equalist'. The combination approach offered in fig. 5 allows a variety to exhibit variation and yet be more central at the same time.

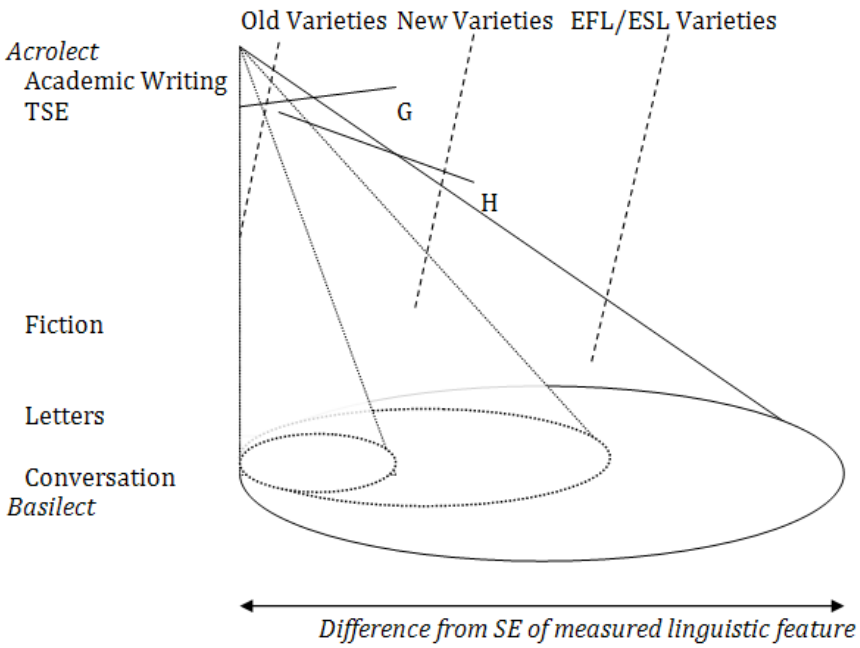

Fig 6.1 Revised Model of World English Varieties

Clearly, any model has its limitations, and the boundaries between varieties must be considered fuzzy; however, the advantage of this hybrid is that it embodies the distinctions and similarities between, on the one hand variety types, and on the other register types. This model assumes a SE and the possibility of 'social chat'. In theory, there is no reason why even an EFL learner may not approach NS-like performance, the left of the model, depending on variables like age and exposure. Unlike the other models, any language, or pragmatic based use, produced by any user of English can be placed precisely on this model and compared to any other, depending on the linguistic feature(s) being assessed, and the domain in which it occurs. Inappropriate register in a domain, e.g. too formal a tone in casual conversation, can be represented for an individual or variety by raising the relative register up the lectal scale on the left. For illustration purposes only, G marks ICE-GB/TSE and H that of ICE-HK/TSE when considering major uses of *the* as described above.

References

Abney, S., (1987). *The English Noun Phrase in its Sentential Aspect*. Doctoral dissertation,

Benson, P. (2000) Hong Kong Words: Variation and Context World Englishes Vol 19/3 Oxford: Blackwell.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999) Longman Grammar of Spoken and Written English. London: Longman.

Bolt, P., & Bolton, K. (1996) International Corpus of English in Hong Kong in Greenbaum, S. (1996) Comparing English World Wide Oxford: Clarendon Press.

Bolton, K. (2005) Where WE Stands: approaches, issues and debate in World English World Englishes Vol. 24/1.

Bolton, K. (2000) The Sociolinguistics of Hong Kong and the Space for Hong Kong English World Englishes Vol. 19/3 Oxford: Blackwell.

Bolton, K. (2002) Chinese Englishes: From Canton jargon to global English World Englishes Vol. 21/2.

Butler, S. (1997) Corpus of English in Southeast Asia: Implications for a regional dictionary in Bautista, M., et al English is and Asian Language: The Philippine context. Manila: The MaQuarie Library.

Bybee, J. (2003). "Cognitive processes in grammaticalization." The new psychology of language (Cognitive and functional approaches to language structure) 2: 145-167.

Christopherson, P. (1939) The Articles: A Study of their Theory and Use in English England: Humphrey Milford: OUP.

Crystal, D. (1995) The Cambridge Encyclopaedia of the English Language UK Cambridge: CUP.

Crystal, D. (2001) The Future of Englishes in Burns, Anne and Coffin, Caroline eds (2001) Analysing English in a Global Context : a reader London : Routledge.

Croft, William (2001) Radical Construction Grammar. UK: Oxford University Press,

Croft, W. (2002). Typology and universals, UK: Cambridge University Press.

Croft, William and Poole, Keith T., (2004) "Inferring Universals from Grammatical Variation: Multidimensional Scaling for Typological Analysis". Available at SSRN: http://ssrn.com/abstract=1154073 accessed 09/06/2010

Gleason, H. (1965) Linguistics and English Grammar New York: Holt, Reinhart and Winston.

Goldberg, A. E. (1995). Constructions: A construction grammar approach to argument structure, USA: University of Chicago Press.

Greenbaum, S. ed. (1996) Comparing English World Wide Oxford: Clarendon Press.

Halliday, M., & Hasan, R. (1976) Cohesion in English. London: Longman Group Ltd.

Halliday, M.,A.,K. (2003) Written Language, Standard Language, Global Language World Englishes Vol. 22/4 Oxford: Blackwell.

Haspelmath, M. (2003). "The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison." The new psychology of language: Cognitive and functional approaches to language structure 2: 211–242.

Hawkins, John, A. (1978) Definiteness and Indefiniteness London: Croom Helm.

Huddleston, R., & Pullum , G.,K. (2002) The Cambridge Grammar of the English Language Cambridge: Cambridge University Press.

Huddleston, R. (1984) Introduction to the Grammar of English. Cambridge: CUP

Hudson, R. (1992) Teaching Grammar Oxford: Blackwell.

Huebner, T. (1983) A Longitudinal Analysis of the Acquisition of English Ann Arbor, Michigan USA: Karoma Press.

Ionin, T., Ko, H., & Wexler, K. (2002) Article Semantics in L2-acquisition: the role of specificity at www.msu.edu/~jk13/Abs.Ionin_Ko_Wexler.pdf.

Jesperson, O., (1933) Essentials of English Grammar London: George Allen % Unwin

Jonson, B. (1640) The English Grammar (reprint 1972) Menston, Yorkshire: Scolar Press.

Kachru, Braj B. (1985) Standards, Codification and Linguistic Realm: The English language in the outer circle in Quirk, R., and Widdowson, H., G. (1985) English in the World Cambridge: CUP.

Kachru, Braj B. (1986) The Alchemy of English. The Spread, Functions and Models of Non-Native Englishes Oxford: Pergamon Press Ltd.

Kachru, Braj B. and Nelson, Cecil (2001) World Englishes in Burns, Anne and Coffin, Caroline eds (2001) Analysing English in a Global Context : a reader London : Routledge.

Langacker, R. (1987). Foundations of Cognitive Grammar, Volume I, Theoretical Prerequisites. Stanford, California: Stanford University Press.

Langacker, R. (2008). Cognitive grammar: A basic introduction. Oxford: Oxford University Press.

Lyons, C., (1999) Definiteness UK: Cambridge: CUP.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985) A Comprehensive Grammar of the English Language. Harlow: Longman Group Press.

Sand, A. (2004) Shared Morpho-Syntactic Features in Contact Varieties of English: article use World Englishes Vol. 23/2.

Spinillo, M. (2005) Reconceptualising the English Determiner Class unpublished PhD Thesis. London: UCL.

Widdowson, H. (1997) EIL, EFL, ESL: Global issues and local interests. World Englishes Vol. 16/1.

Yanu, Y (2001) World Englishes in 2000 and Beyond World Englishes Vol. 20/2.

Yow, S. (2001) Guangdong to Trail English as Medium in South China Morning Post October 20th 2002.

Yuen-Ying, Chan (2000) The English Language Media in Hong Kong World Englishes Vol. 9/3 Oxford: Blackwell.