World Englishes, Vol. 25, No. 1, pp. 115–129, 2006. 0883–2919

RESEARCH REPORTS

The core and periphery of world Englishes: a corpus-based exploration

GERALD NELSON*

ABSTRACT: This paper is concerned with the 'common core' of world Englishes. This is defined as the set of features and characteristics which all varieties have in common. It is also concerned with the 'periphery', that is, those features which are unique to individual varieties, and are not shared with any other variety. The study is based on comparisons of wordlists from electronic corpora of six varieties, and it attempts to examine the core and periphery of these in both quantitative and qualitative terms. It offers a measure of the relative sizes of the core and periphery, and goes on to compare the core and periphery in terms of their grammar and morphology. The core is found to be grammatically conditioned to a much greater degree than the periphery, which in turn results in very striking morphological differences between the two.

INTRODUCTION

For Quirk *et al.* (1985: 16), the existence of a common core of world Englishes is an undisputed fact:

> A common core or nucleus is present in all the varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others. It is this fact that justifies the application of the name 'English' to all the varieties.

Intuitively, it is perfectly reasonable to believe that such a core really does exist, if only because the varieties are mutually intelligible. However, intuition alone can tell us very little about the dimensions of the core, or how it is constituted. In its simplest form, we might visualize the common core in terms of a Venn diagram (Figure 1), where each variety of English is imagined as a circle, and the core is that area where all the circles overlap with each other. It follows, too, that there must also be a periphery of world Englishes, that is, an area where no overlap takes place. This will consist of items and characteristics which are distinctive of one variety only.

The aim of this paper is to examine the core of world Englishes both quantitatively and qualitatively, using parallel corpora as the basis of study. It aims to discover (a) the dimensions of the core (exactly how much overlap exists among the varieties?), and (b) the nature of the core (what is the 'set of grammatical and other characteristics' that makes up the core?). Answering these questions will clearly be of both theoretical and practical interest. In practical terms, the size and content of the core will be of interest to lexicographers of world Englishes, EFL teachers, and authors of EFL textbooks.

* Department of English Language and Literature, University College London, Gower St, London WC1E 6BT, UK. E-mail: g.nelson@ucl.ac.uk

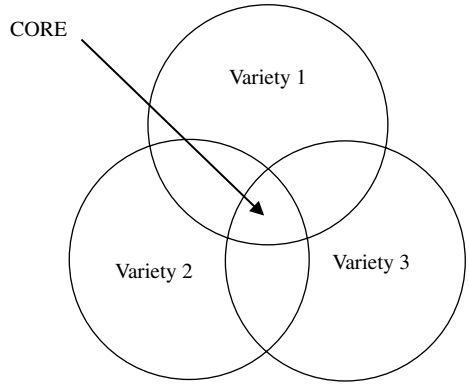

Figure 1. Visualization of the core

DATA

This paper will examine the core (and the periphery) of world Englishes by studying six electronic corpora from the International Corpus of English (ICE: Greenbaum, 1996). The ICE project was initiated in 1990 in order to provide parallel corpora of varieties of English from around the world, to be used as the basis of cross-variety comparisons. In the present study I examine six of the currently available ICE corpora: Great Britain (ICE-GB), New Zealand (ICE-NZ), India (ICE-IND), Singapore (ICE-SIN), the Philippines (ICE-PHI), and Hong Kong (ICE-HK).[1] The corpora consist of one million words each, sampled across the same 32 text categories (both written and spoken), and all samples date from 1990 or later (Nelson, 1996). At the time of writing, the corpora exist in lexical form only, with the exception of ICE-GB, which has been POS-tagged and parsed (Nelson, Wallis, and Aarts, 2002). Because of this, the comparisons made between the corpora in this paper will be largely based on wordlists, so we will be looking chiefly at the 'lexical' core and the 'lexical' periphery. However, as we will see, some grammatical information can be derived from the study of wordlists, since lexis is ultimately grammatical. We shall also use the wordlists to examine the morphology of the core and of the periphery.

PROCEDURE

Each of the six corpora under review was indexed using TACT,[2] and an alphabetical wordlist was produced for each corpus. Typically, a corpus of one million words (tokens) consists of around 50,000 different words (types). In order to make the lists manageable in terms of size, I have excluded from them all cardinal numerals, such as *1*, *10*, *100*, as well as hyphenated words which include numerals as part of the compound, such as *5-mile*, *10-gallon*, and *24-hour*. This has the effect of reducing each of the wordlists by about 10,000 items, so the resulting wordlists contain around 40,000 types each. The exact figures are shown in Table 1, which also shows the top and bottom 25 words in each of the six corpora.

As Table 1 shows (not surprisingly), the indefinite article *a* is common to all the varieties, so it is clearly part of the core. At the other end of the alphabetical lists, the item *zoological* is also found in all six wordlists, so this too may be called a 'core' lexical

Table 1. Top and bottom 25 items on each wordlist

Great Britain	New Zealand	India	Singapore	Philippines	Hong Kong
a	*a*	*a*	*a*	*a*	*a*
a-level	*a+*	*a-days*	*aa*	*a&m*	*a′*
a-rev	*a−*	*a-n-a-n-d*	*aaa*	*a + b*	*a′s*
a-scan	*a − b*	*a-ring*	*aag*	*a-h-hs*	*a + s*
a-scans	*a-half*	*a − t*	*aah*	*a-ok*	*a-*
a-wooing	*a/*	*a-team*	*aaia*	*a-okay*	*a-b*
a-z	*a/a +*	*a/c*	*ab*	*a-plenty*	*a-level*
a/b	*a/b*	*aa*	*aback*	*a-visiting*	*a-levels*
a/cm	*a/d*	*aaa*	*abacus*	*aa*	*a-listers*
aa	*aa*	*aac*	*abandon*	*aaahhh*	*a-mee'*
aaah	*aaaagh*	*aad*	*abandoned*	*aacc*	*a/c*
aac	*aaah*	*aadam*	*abandoning*	*aaecp*	*aa*
aah	*aahed*	*aadat*	*abandonment*	*aalis*	*aand*
aand	*aano*	*aadhar*	*abasad*	*aalisan*	*aare*
aar	*aaqib*	*aadmi*	*abbas*	*aaron*	*aaron*
aargh	*aargh*	*aage*	*abbott*	*aawitan*	*ab*
aaron	*aaron*	*aahariya*	*abbotti*	*ab*	*abab*
aaronite	*ab*	*aai*	*abbrevations*	*aba*	*abaca*
aaronites	*ababa*	*aaichi*	*abbrevia*	*abac*	*aback*
aarts	*abandoment*	*aaichich*	*abbreviated*	*abaca*	*abandon*
aat	*abandon*	*aaina*	*abbreviations*	*abacus*	*abandoned*
ab	*abandoned*	*aaj*	*abc*	*abad*	*abandoning*
abab	*abandoniment*	*aajeev*	*abcs*	*abadia*	*abandons*
abacus	*abandoning*	*aakal*	*abd*	*abadilla*	*abate*
abajo	*abandonment*	*aakals*	*abdomen*	*abalone*	*abatement*
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
zon	*zip-fasteners*	*zise*	*zinc*	*zip*	*zinc*
zonal	*zipped*	*zn*	*zip-code*	*zipped*	*zip*
zonation	*zippered*	*zona*	*zoological*	*zm*	*zipper*
zone	*zither*	*zonal*	*zoology*	*zobel*	*zire*
zones	*zn*	*zone*	*zoom*	*zobels*	*ziyang*
zonfrillo	*zoellner*	*zones*	*zooming*	*zoilo*	*zodiac*
zoo	*zoisite*	*zonewise*	*zoos*	*zolu*	*zoe*
zoological	*zonation*	*zoo*	*zoran*	*zombies*	*zone*
zoologists	*zone*	*zool*	*zorun*	*zone*	*zoned*
zooming	*zoned*	*zoological*	*zou*	*zones*	*zones*
zooplankton	*zones*	*zoologist*	*zubir*	*zoning*	*zoo*
zoos	*zoning*	*zoologists*	*zubit*	*zoo*	*zoological*
zoox	*zonked*	*zoology*	*zubri*	*zoological*	*zoology*
zortech	*zonulae*	*zoom*	*zukifli*	*zoology*	*zoom*
zosimus	*zoo*	*zoomed*	*zul*	*zoom*	*zoomer*
zox	*zool*	*zooming*	*zulkifi*	*zooming*	*zooming*
zoxon	*zoological*	*zooms*	*zulkifli*	*zooms*	*zoos*
zoxtia	*zoology*	*zoos*	*zulklifi*	*zoonosis*	*zubin*
zr-	*zoom*	*zoroastrian*	*zulu*	*zooplanktonic*	*zuhai*

(Table 1, continued)

Great Britain	New Zealand	India	Singapore	Philippines	Hong Kong
zulu	*zoomed*	*zoroastrianism*	*zulus*	*zubay*	*zuni*
zux	*zooming*	*zubel*	*zuraidah*	*zubiri*	*zunxin*
zuxon	*zoos*	*zubin*	*zuran*	*zubur*	*zuo′er*
zworykin	*zulu*	*zubovsky*	*zurax*	*zulueta*	*zurich*
zyx	*zum*	*zulu*	*zurich*	*zurich*	*zvereva*
zyxton	*zz*	*zuzana*	*zvezda*	*zurita*	*zvs*
Types = 41,086	**39,366**	**41,330**	**34,367**	**39,536**	**34,125**

item, at least as far as these six varieties are concerned. At the other extreme, the item *zon* is found only in ICE-GB, and is therefore peripheral. Between these extremes we have items like *zoo*, which occur in five of the six corpora. This is clearly less peripheral than *zon*, but it still does not qualify for membership of the core, since it is missing from one of the corpora. There is, of course, a gradient from the core to the periphery, and for this reason I shall use the term *absolute core* to refer to items which occur in all six corpora, and *absolute periphery* to refer to items which occur in one corpus only. In terms of our six corpora, we can visualize this in terms of the Venn diagram shown in Figure 2.

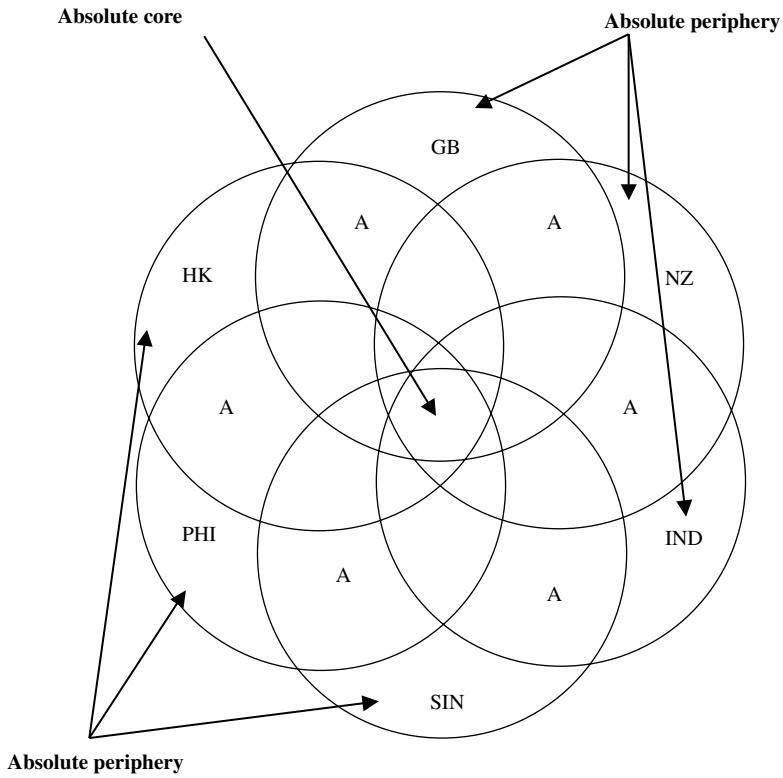

Figure 2. Visualization of the absolute core and the absolute periphery

The next step is to calculate the extent of the overlap between each corpus and every other corpus (the six areas marked 'A' in Figure 2). To do this, the wordlists were combined in pairs (ICE-GB with ICE-NZ, ICE-GB with ICE-IND, and so on). The results of this are shown in Table 2.

By combining the wordlists from ICE-GB and ICE-NZ, a new wordlist was produced, consisting of 58,825 types (C = 58,825 in Table 2). This combined wordlist was then indexed, in order to calculate the extent of the overlap (O) between the two original lists. Items with a frequency of two in the combined wordlist clearly derive from both corpora (O = 21,599 in Table 2). This figure is then shown as a percentage of the total (in this case, 36.7%). This procedure was then carried out for each pair of wordlists in the study. Table 2 shows that the extent of the overlap between the varieties is fairly consistent, at around 30 per cent. What this means is that the varieties share about 30 per cent of types with every other variety. The figure for ICE-India's overlap with other varieties is slightly lower than average, and probably worth further investigation, though in statistical terms the figure is not significant.

THE SIZE OF THE CORE

Table 2 shows the extent to which the individual corpora overlap with each other (in terms of types), but it does not indicate the size of the core, since it compares pairs of corpora, rather than all six. To calculate this, we must combine all six wordlists into a single wordlist, and then index this new list. In the new list, items with a frequency of six will be those which occur in all six of our corpora, and may therefore be said to constitute the 'absolute core' of the corpora under review. At the other end of the scale, items with a

Table 2. Extent of overlap between individual wordlists (C = Number of types in combined lists; O = Overlap, i.e., number of types common to both wordlists)

	ICE-GB	ICE-NZ	ICE-IND	ICE-SIN	ICE-PHI	ICE-HK
ICE-GB	–					
ICE-NZ	C = 58,825 O = 21,599 =**36.7%**	–				
ICE-IND	C = 62,986 O = 19,402 =**30.8%**	C = 61,621 O = 19,049 =**30.9%**	–			
ICE-SIN	C = 56,100 O = 19,321 =**34.4%**	C = 54,384 O = 19,319 =**35.5%**	C = 57,499 O = 18,168 =**31.6%**	–		
ICE-PHI	C = 60,784 O = 19,807 =**32.6%**	C = 59,207 O = 19,666 =**33.2%**	C = 62,335 O = 18,502 =**29.7%**	C = 55,056 O = 18,814 =**34.2%**	–	
ICE-HK	C = 55,857 O = 19,326 =**34.6%**	C = 54,416 O = 19,048 =**35.0%**	C = 57,747 O = 17,684 =**30.6%**	C = 49,876 O = 18,588 =**37.3%**	C = 54,881 O = 18,753 =**34.2%**	–
Average overlap	**33.8%**	**34.3%**	**30.7%**	**34.6%**	**32.8%**	**34.3%**

Table 3. All six corpora compared: top and botton 25 items from each 'overlap' list

Items in all 6 corpora **absolute core**	Items in 5 corpora	Items in 4 corpora	Items in 3 corpora	Items in 2 corpora	Items in 1 corpus only **absolute periphery**
a	*abandonment*	*aaron*	*a-b*	*aaa*	*aaaagh*
aa	*abandons*	*abbas*	*aback*	*aaah*	*aaahhh*
ab	*abbreviated*	*abbey*	*abacus*	*aac*	*aacc*
abandon	*abducted*	*abbreviation*	*abate*	*aah*	*aad*
abandoned	*abiding*	*abdul*	*abating*	*aahed*	*aadam*
abandoning	*abject*	*abel*	*abattoir*	*aargh*	*aadat*
abdomen	*abnormally*	*aberdeen*	*abba*	*aba*	*aadhar*
abdominal	*abode*	*aberration*	*abbott*	*abab*	*aadmi*
abide	*abolish*	*aberrations*	*abbreviations*	*abaca*	*aaecp*
abilities	*abortions*	*ably*	*abduction*	*aban*	*aag*
ability	*abraham*	*abnormalities*	*abdullah*	*abatement*	*aage*
able	*abreast*	*abnormality*	*abeyance*	*abb*	*aahariya*
abnormal	*abstain*	*abolishing*	*abhorred*	*abby*	*aai*
aboard	*abstraction*	*abort*	*abided*	*abc*	*aaia*
abolished	*acceleration*	*abounds*	*able-bodied*	*abcd*	*aaichi*
abolition	*accented*	*abrasion*	*abo*	*abcs*	*aaichich*
aborted	*acceptability*	*abruptness*	*aboriginal*	*abd*	*aaina*
abortion	*accessed*	*abs*	*abortive*	*abdicated*	*aaj*
abound	*accessing*	*abscess*	*abou*	*abdication*	*aajeev*
about	*accessory*	*absorber*	*above-average*	*abdomens*	*aakal*
above	*accommodated*	*absorbing*	*abrasive*	*abduct*	*aakals*
abroad	*accommodating*	*abstaining*	*abrogate*	*abductor*	*aala*
abrupt	*accomodation*	*abstracts*	*absences*	*abe*	*aale*
abruptly	*accomplishment*	*abundantly*	*abso*	*aber*	*aalis*
absence	*accomplishments*	*abuser*	*absolution*	*abet*	*aalisan*
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
yield	*write-up*	*yearns*	*yuan*	*zestful*	*zukifli*
yields	*wrongdoing*	*yeast*	*yuck*	*zeus*	*zul*
york	*wrongful*	*yell*	*yugoslavian*	*zhan*	*zulkifi*
you	*wrongs*	*yelled*	*yuh*	*zhao*	*zulkifli*
young	*ya*	*yells*	*yummy*	*zhen*	*zulklifi*
younger	*yearly*	*yep*	*yun*	*zheng*	*zulueta*
youngest	*yearning*	*yew*	*zag*	*zhu*	*zulus*
youngsters	*year-round*	*yi*	*zealanders*	*zhuang*	*zum*
your	*yellowish*	*yoga*	*zealous*	*zi*	*zuni*
yours	*yen*	*yoghurt*	*zed*	*zigzag*	*zunxin*
yourself	*yielded*	*yolk*	*zee*	*zigzagging*	*zuo'er*
yourselves	*yielding*	*yong*	*zen*	*zigzags*	*zuraidah*
youth	*yo*	*yorkshire*	*zenith*	*zimmerman*	*zuran*
youthful	*yoke*	*yrs*	*zest*	*zing*	*zurax*
youths	*youngster*	*yuk*	*zhang*	*zipping*	*zurita*
yugoslav	*yr*	*yung*	*zhi*	*zn*	*zux*
yugoslavia	*yu*	*yuppies*	*zhou*	*zodiac*	*zuxon*

(Table 3, continued)

Items in all 6 corpora **absolute core**	Items in 5 corpora	Items in 4 corpora	Items in 3 corpora	Items in 2 corpora	Items in 1 corpus only **absolute periphery**
zealand	*yup*	*yuri*	*zia*	*zoe*	*zuzana*
zero	*yuppie*	*zaire*	*zig-zag*	*zoned*	*zvereva*
zinc	*yvonne*	*zeal*	*zilch*	*zoning*	*zvezda*
zone	*z*	*zebra*	*zimbabwe*	*zool*	*zvs*
zones	*zip*	*zeros*	*zipped*	*zoologists*	*zworykin*
zoo	*zoology*	*zero-sum*	*zonal*	*zoomed*	*zyx*
zoological	*zoom*	*zig*	*zonation*	*zooms*	*zyxton*
zooming	*zoos*	*zulu*	*zurich*	*zubin*	*zz*
Types = 11,653 = 11%	**4,790 = 4%**	**5,106 = 5%**	**7,074 = 6%**	**12,547 = 11%**	**69,149 = 63%**

frequency of one derive from just one corpus. These constitute the 'absolute periphery'. Table 3 shows the top and bottom 25 items in each of these 'overlap' lists, that is, the top and bottom 25 items which occur in all six corpora, in five corpora, in four corpora, and so on.

The combined list of six individual wordlists consists of 110,319 types. Of these, 11,653 (11%) are found in all six wordlists. So the 'absolute core' represents just 11 per cent of types. At the other extreme, the absolute periphery – items occurring in one corpus only – consists of 69,149 types, or 63 per cent of all types.

In terms of types, therefore, the absolute core is quite small, and the absolute periphery is very large indeed, at well over half of all types. However, Table 3 tells us nothing about the frequency of the items at the core and at the periphery, since it is based only on types, not on tokens. It is reasonable to suppose that items in the absolute core – such as *a*, *able*, and *about* – occur with much greater frequency than items such as *aaaagh* and *aaahhh*, which belong to the absolute periphery. Although there are far fewer core items, they are likely to have much greater frequency than peripheral items. To investigate this further, it is clearly not practical to examine the frequencies of all 11,653 absolute core items, much less those of the 69,149 items in the absolute periphery. Instead, I have taken a random sample of 50 items from each list and calculated their total frequencies in the corpora.[3] Table 4 shows the 50 words selected at random from each of the lists.

Using the 50 random words from each list, we can calculate the average frequency of each word, and therefore their overall frequency, that is, the total number of tokens represented by the absolute core and by the absolute periphery, and by everything in between. These figures are shown in Table 5.

Table 5 highlights, above all, the vast difference in frequency between items in the core and items in the periphery. Items in the absolute core occur on average 454.8 times, whereas items in the absolute periphery occur just 2.1 times. The frequencies decline steadily as we move out from the absolute core to the absolute periphery, though by far the sharpest decline occurs as we move from the absolute core (454.8) to the next column (items in five corpora, with average frequency 27.4). This indicates that in terms of frequencies, there is a very radical difference between the absolute core and everything else, and appears to justify our delimiting the absolute core to the area where all six corpora overlap. The absolute core represents 91 per cent of

Table 4. Fifty words selected at random from each wordlist

6 corpora	5 corpora	4 corpora	3 corpora	2 corpora	1 corpus
apply	*accessed*	*allusions*	*abacus*	*ances*	*aad*
break	*acquaintances*	*analyzed*	*algeria*	*assam*	*alkylated*
constituency	*airliner*	*attire*	*amazonian*	*asymmetry*	*alow*
continents	*airports*	*barnes*	*amo*	*attorneys*	*apneko*
cords	*alleging*	*blockage*	*anniversaries*	*bogeyman*	*ballyhooed*
corridor	*ambush*	*candy*	*asexual*	*borland*	*balt*
denote	*assays*	*casey*	*asteroid*	*bounding*	*barsch*
disappointed	*avenues*	*celebrates*	*aye*	*cachet*	*blockwork*
dough	*bliss*	*chilli*	*bookkeeping*	*crochet*	*charise@thomasians.com*
earlier	*boxing*	*chronicles*	*cavalry*	*delighting*	*choy-yin*
exemption	*contemporaries*	*clambered*	*communicators*	*depa*	*clrc*
finds	*craving*	*commercialisation*	*complicity*	*duty-free*	*comming*
floods	*cylinders*	*cora*	*concocting*	*elongate*	*compacts*
hand	*deductions*	*cringe*	*coventry*	*fantastical*	*congresses*
identifying	*deserts*	*dens*	*diode*	*fielder*	*falacca*
idiot	*disrespectful*	*dependable*	*dreary*	*gec*	*gesundheit*
imposing	*edgar*	*dictation*	*forceps*	*goyang*	*gossypol*
increasing	*endorses*	*directional*	*forfeit*	*harpooned*	*grockle*
incubation	*engagements*	*dylan*	*haiti*	*high-spirited*	*hihihihihihi*
ironically	*enigmatic*	*east-west*	*horrid*	*hot-water*	*hutcheson*
jails	*excludes*	*emigrated*	*hotdogs*	*hypothyroidism*	*iyong*
kick	*fullest*	*entitlements*	*indifferently*	*imbalances*	*leebov*
knife	*gratefully*	*forecasting*	*inter-tidal*	*inferno*	*leukodepletion*
massacre	*homogenous*	*frenchman*	*jacked*	*inflamation*	*lugi*
may	*idealism*	*gossipy*	*japs*	*instal*	*lully*
personnel	*inception*	*gypsies*	*manipulates*	*knackered*	*massi*
poem	*longed*	*johannesburg*	*montgomery*	*maja*	*mathmaticly*
positively	*makeshift*	*languishing*	*navratilova*	*migrations*	*mid-chest*
proposed	*med*	*liam*	*non-profit*	*minted*	*motivator*
punished	*migrants*	*liner*	*overdrive*	*muni*	*munsifs*
sank	*mop*	*marion*	*pathologies*	*pertained*	*non-graphical*
senior	*murphy*	*normalised*	*polluter*	*phenotypes*	*non-kin*
signature	*narrowly*	*octo*	*protoplasm*	*pre-set*	*numinous*
sons	*nocturnal*	*paradigms*	*quarter-final*	*pro-beijing*	*petalled*
storms	*pedal*	*pathogens*	*ransacking*	*prying*	*post-box*
submission	*prohibiting*	*prudential*	*recompense*	*readjustment*	*premanance*
substantial	*queried*	*pulsating*	*rectification*	*robyn*	*qt*
surely	*rotting*	*pylons*	*refurbished*	*roommates*	*rolf*
synthetic	*rs*	*recyclable*	*reiterates*	*rosyth*	*rubbishy*
tension	*satire*	*sch*	*sayings*	*sawyer*	*santillan*
testified	*self-esteem*	*sheltering*	*seminary*	*silage*	*sarma*
themes	*shameful*	*smog*	*shepherding*	*top-quality*	*sleekest*
tied	*spawned*	*soundness*	*shorn*	*top-selling*	*sterner*
trailing	*taxable*	*stephanie*	*sourcing*	*tradesmen*	*stockyard*
turn	*taxpayer*	*trot*	*taint*	*triennial*	*thoselimitations*
uniqueness	*towers*	*undecided*	*throats*	*unabashedly*	*tranquillisers*
waves	*walkway*	*wallpaper*	*turnaround*	*uselessness*	*underpasses*
weeks	*wavy*	*workmanship*	*versatility*	*waster*	*unreported*
wife	*weighed*	*yep*	*whiskers*	*wcs*	*wein*
wonderful	*wrecked*	*zaire*	*yawn*	*wisma*	*zum*

Table 5. Total frequencies of tokens, based on random 50 words from each list

	6 corpora	5 corpora	4 corpora	3 corpora	2 corpora	1 corpus	Totals
A: Total freq. of random 50 words	22,738	1,373	1,013	566	263	106	
B: Average freq. of each word (A/50)	**454.8**	**27.4**	**20.3**	**11.3**	**5.3**	**2.1**	
C: Total no. of types	11,653	4,790	5,106	7,074	12,547	69,149	110,319 types
Overall frequency (B × C)	5,299,784	131,246	103,651	79,936	66,499	146,595	5,827,711 tokens
% of all tokens	**91%**	**2%**	**2%**	**1%**	**1%**	**3%**	**100%**
% of all types	**11%**	**4%**	**5%**	**6%**	**11%**	**63%**	**100%**

all the tokens in the six corpora (5.2 million words out of a total of 5.8 million). The absolute periphery, by contrast, represents only 3 per cent of all tokens in the corpora.

We are now in a position to describe the dimensions of the core and of the periphery. The absolute core consists of relatively few lexical items (11,653 in total, or 11% of all types), but these occur with very high frequency (average frequency 454.8). In contrast, the absolute periphery consists of very many lexical items (69,149, or 63% of types), which occur very infrequently (average frequency of 2.1).

Having established the dimensions of the core and periphery in quantitative terms, the remainder of this paper will examine these areas qualitatively, looking at the kinds of words that occur in the absolute core and in the absolute periphery.

WORD LENGTH AT THE CORE AND AT THE PERIPHERY

Figure 3 shows a graph of average word length in each of the six wordlists, as we move from the absolute core to the absolute periphery. It shows a very steady increase in word length as we move out from the absolute core. Items in the absolute core have an average length of 7.3 characters, compared with those at the absolute periphery, which have an average length of 8.3 characters. The result is not at all surprising: it simply bears out Zipf's Law regarding the inverse relationship between word length and word frequency.

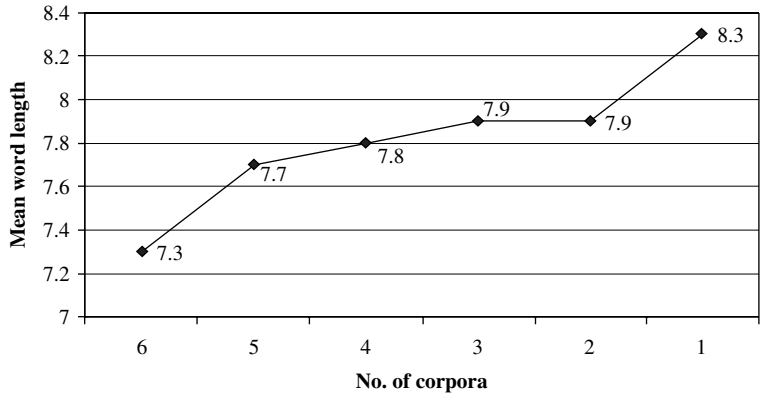

Figure 3. Variation in word length from core to periphery

WORD CLASSES IN THE CORE AND IN THE PERIPHERY

On the basis of the 50 random words from the absolute core, and the 50 from the absolute periphery (see Table 4), we can examine the individual words in the contexts in which they occur in the corpora, to determine what word classes they belong to. Table 6 shows the breakdown by word class of the 50 random words from the absolute core.

They consist of the central word classes: nouns, verbs, adjectives, and adverbs. In contrast with this, the 50 words from the absolute periphery are much more diverse (Table 7). The vast majority are nouns – including proper nouns, which do not feature in the absolute core – and there are no verbs. The remainder are a mixture of misspellings, foreign and indigenous words, indeterminate forms, and nonce formations. Interestingly, all six varieties under review contribute to the absolute periphery, as well as (by definition) to the absolute core.

MORPHOLOGY OF THE CORE AND OF THE PERIPHERY

Even a cursory glance at the wordlists in Table 1 reveals that items in the core and in the periphery are morphologically very distinct. In general, the absolute core contains items with regular English suffixes, such as *ability*, *abortion*, and *abruptly*, whereas the absolute periphery contains items which are morphologically very irregular, such as *aacc, aag*, and *aala*. Using the *Linguistica* program (Goldsmith, 2001), it is possible to carry out a fairly comprehensive analysis of suffixation in all the items on the wordlists.[4] The results of this analysis are shown in Table 8, where the most frequent regular suffixes are shown for both the absolute core and the absolute periphery.

In terms of suffixation, the contrast between core and periphery is very striking. The absolute periphery contains very few regular suffixes, and these are mostly noun-forming suffixes. The absolute core, on the other hand, displays a very wide range of suffixes associated with a very wide range of word classes. These results corroborate our earlier

Table 6. Word classes in the absolute core, based on 50 random words

Wordforms	Classification	Number
constituency, continents, cords, corridor, dough, exemption, idiot, incubation, jails, knife, personnel, poem, signature, sons, storms, submission, tension, themes, uniqueness, weeks, wife	Common nouns	21
apply, break, denote, finds, floods, hand, identifying, imposing, increasing, kick, massacre, proposed, punished, sank, testified, tied, trailing, turn, waves	Verbs (or nouns)	19
disappointed, synthetic, senior, substantial, wonderful	Adjectives or participial adjectives	5
earlier, ironically, positively, surely	Adverbs	4
may	Modal verbs	1

Table 7. Word classes in the absolute periphery, based on 50 random words

Wordforms and source corpus	Classification	Number
compacts, stockyard (NZ) *motivator* (PHI) *mid-chest, non-kin, underpasses* (SIN) *gossypol, munsifs* (IND) *blockwork, leukodepletion* (HK) *grockle, post-box* (GB)	Common nouns	12
Balt, Congresses, Rolf, Sarma (IND) *Falacca* (GB) *CLRC* (SIN) *Leebov, Santillan* (PHI) *Hutcheson* (NZ) *Choy-yin, Lully* (HK)	Proper nouns	11
numinous, rubbishy (GB) *alkylated, ballyhooed, sterner* (PHI) *petalled* (SIN) *non-graphical* (NZ) *sleekest, unreported* (HK)	Adjectives or participial adjectives	9
mathmaticly, *tranqullisers* (NZ) *premanance* (SIN) *comming, thoselimitations* (GB)	Misspellings	5
barsch, Gesundheit, wein, zum (NZ) *Massi* (GB)	Foreign words	5
iyong, lugi (PHI) *aad, apneko* (IND)	Indigenous words	4
alow (GB) *qt* (NZ) *charise@thomasians.com* (PHI)	Indeterminate/incomplete/email	3
hihihihihihi (PHI)	Nonce words	1

findings about the distribution of word classes. They show clearly that the core of world Englishes is grammatically conditioned to a much greater extent than the periphery.

No such striking contrast is found when we look at prefixation. Intuitively, we might have expected that the absolute periphery would contain 'technical' or 'scientific' vocabulary, the kind used frequently in specialized contexts, and very rarely in 'core' contexts. So we might expect items with typically 'technical' Greek and Latin prefixes, such as *astro-*, *micro-*, and *tele-*, to predominate at the absolute periphery. As Table 9 shows, this is certainly not the case.

The results shown in Table 9 were produced by searching the absolute core and the absolute periphery wordlists for wordforms with typically 'technical' or 'scientific' prefixes. The following 50 prefixes were counted in each list:

aero- ante- anthrop- astro- audio- auto- bio- cardi- cent- circum- contra- crypto- deca- dys- eco- geo- haem- hetero- homeo- homo- hydr- infra- inter- kilo- mal- matri- mega- micro- mille- mini- mono- multi- neo- omni- para- photo- physio- proto- pseudo- psycho- quasi- retro- semi- socio- super- techn- tele- theo- thermo- ultra-[5]

Table 8. Morphology of the absolute core and of the absolute periphery: suffixes

Absolute core: 11,653 types		Absolute periphery: 69,149 types	
Suffix	No.	Suffix	No.
-ed	1,926	*-es*	426
-ing	1,900	*-ed*	392
-s	1,714	*-ing*	312
-es	998	*-er*	226
-tion	696	*-al*	218
-ly	596	*-ation*	88
-y	594	*-ity*	64
-er	576		
-al	454		
-ity	244		
-ic	216		
-able	156		
-ous	150		
-ment	140		
-ive	130		
-ar	126		
-est	112		
-ical	102		
-ence	102		
-ness	74		
-ful	72		
-ist	64		
-ture	64		
-ish	54		
-ism	46		
-ics	34		
-ship	32		

Table 9 shows clearly that, whatever else it may contain (and it is very diverse, as we have seen), the absolute periphery is not specifically the home of technical, scientific, or specialized vocabulary. Just 3 per cent of its vocabulary is of this type, only slightly more than is found at the absolute core (2.1%).

VOCABULARY PROFILING

Up to now, we have derived results based solely on the six ICE corpora under review. Clearly, as more ICE corpora become available in the near future, these can be examined in

Table 9. Morphology of the absolute core and of the absolute periphery: prefixes

	No. of types with a 'technical'/'scientific' prefix	Total no. of types	%
Absolute core	241	11,653	2.1%
Absolute periphery	2,086	69,149	3.0%

the same way, an effort which will doubtless produce more robust results. On the other hand, no matter how many new corpora become available, it is appropriate, also, to compare our corpus-based results with independently-produced wordlists from outside the ICE project.

One way of doing this is to use Vocabulary Profiling, which compares our wordlists with independently derived wordlists, including West's *General Service List of English Words* (West, 1953) and Coxhead's *The Academic Wordlist* (Coxhead, 2000; Nation and Bauer, 1993). Paul Nation's *VocabProfile* program[6] compares wordlists with three standard 'base' lists derived from these and other standard lists. These are:

- Base List 1: the 1,000 most frequent words in English,
- Base List 2: the second 1,000 most frequent words in English,
- Base List 3: words not found in the first 2,000 words (not on Base Lists 1 or 2), but which are frequent in upper secondary school texts from a wide range of subjects.

The program was used to compare our absolute core and absolute periphery lists with the three base lists, and the results are shown in Table 10.

Once again, the contrast between core and periphery is quite striking: almost 99 per cent of the vocabulary at the periphery is not found on any of the three base lists. Perhaps the most surprising result is that only 27.7 per cent of absolute core items are found on Base List 1, the most frequent 1,000 words in English. If we include Base List 2, this figure rises to just 45.4% per cent. What does this tell us about the absolute core, or, indeed, about the base lists? The low figure is partly due to the fact that we are not really comparing like with like. The wordlists used in this study are unlemmatized lists, whereas the base lists used in *VocabProfile* are lists of headwords, that is, of uninflected base forms. I will address the issue of lemmatization in the next section. The disparity between the absolute core and the base lists is also due to the fact that the base lists are not derived from world Englishes, but largely from American English and British English. In contrast, the core being examined here is drawn from six varieties, some of which – Hong Kong, Philippines, and Singapore, in particular – very likely introduce new factors into the equation. Further, although West's *General Service List* (1953) has been very influential for many

Table 10. Vocabulary profiles of the absolute core and of the absolute periphery

Absolute core: all six corpora

Word list	Types	%
Base List 1	3,232	27.7
Base List 2	2,057	17.7
Base List 3	1,808	15.5
Not on any list	4,554	**39.1**

Absolute periphery: 1 corpus only

Word list	Types	%
Base List 1	442	0.6
Base List 2	210	0.3
Base List 3	180	0.3
Not on any list	68,649	**98.8**

years, I would suggest that it is now rather old. In the light of the emerging new Englishes, perhaps the time has come to draw up a '*General Service List of World English Words*'.

A NOTE ON METHODOLOGY

As mentioned above, this study is based entirely on unlemmatized wordlists. That is, all the inflectional variants of nouns, verbs, adjectives, and adverbs are treated as distinct forms, so *zoo* and *zoos*, for instance, are treated separately. If the lists were lemmatized, these variants would appear as the base form *zoo*, and so Table 1, for example, would look somewhat different. In fact, *zoo* would then appear in all six lists, and would therefore be part of the absolute core. It is difficult to estimate how much lemmatization of the entire wordlists would alter the overall results presented here, but ideally, in any future study of this topic, lemmatized wordlists should be used in calculating the size of the core and of the periphery. That task will be considerably easier once all of the corpora have been POS-tagged.

On the other hand, if the wordlists were lemmatized, we would not have been able to examine the grammar and morphology of the core and of the periphery. Lemmatization collapses grammatical and morphological distinctions by removing inflections and reducing each item to its base form. So forms of the noun *bank* (*bank*, *banks*), together with forms of the verb *bank* (*bank*, *banks*, *banked*, *banking*) would all appear as the base form *bank*. In doing so, we lose the distinction between the noun and the verb, and many of the suffixes which were used in the morphological analysis. To examine the core and periphery in qualitative terms, the use of unlemmatized lists seems appropriate, though once again, greater accuracy could be achieved using POS-tagged corpora.

CONCLUSION

This paper has afforded a glimpse, albeit a fairly narrow one, of the core and periphery of world Englishes. It is narrow for at least two reasons. Firstly, the six varieties of English under review are by no means intended to be representative of world Englishes generally. As more ICE corpora become available, it will be interesting to see if their inclusion will substantially affect the overall results presented here. Perhaps the inclusion of some African and Caribbean varieties would do so, but this is speculation.

Secondly, the study of the grammar of the core and periphery presented here is based on just 50 random words. Ideally, more words might have been chosen, but when dealing with wordlists of such length, some practical cut-off point has to be chosen. A random set of 500 words, for instance, would certainly have yielded a more comprehensive view of the grammar, but, again, broadening the study in that way would not necessarily alter the overall results to a significant degree.

In summary, these overall results are: (1) the absolute core is very small, in terms of wordforms – just 11 per cent of all types belong to the absolute core, while 63 per cent belong to the absolute periphery. (2) the absolute core is very large, in terms of frequency. It represents 91 per cent of all tokens, while the absolute periphery represents just 3 per cent of tokens.

In short, the absolute core consists of a very small number of lexical items which occur very frequently. The absolute periphery is the converse of this: it consists of a very large number of items which occur very rarely.

This study has also shown:

1. The absolute core contains the central word classes: common nouns, verbs (including modals), adjectives, and adverbs.
2. The absolute periphery chiefly contains proper nouns and common nouns, as well as some adjectives. The remainder is a very diverse set of items, including foreign and indigenous words, misspellings, and nonce words.
3. The absolute core contains a very comprehensive range of standard suffixes, very few of which are to be found in the absolute periphery.
4. There is little difference between core and periphery in terms of Greek and Latin prefixes, that is, the periphery does not contain 'technical' vocabulary to any significant degree.
5. All the varieties under review contribute to the absolute periphery.

In future research, it would be interesting to investigate the extent to which individual varieties contribute to the absolute periphery, and the nature of those contributions, in terms of grammar, morphology, and semantics. More crucially, perhaps, the contributions which individual varieties make to the absolute core remains a fruitful area for further investigation.

NOTES

1. For details of the availability of ICE corpora, see the ICE website, at http://www.ucl.ac.uk/english-usage/ice/index.html. The ICE Hong Kong corpus will be made available to researchers from January 2006. The Hong Kong corpus was funded by the Hong Kong Research Grants Council (RGC), Grant nos HKU 7174/00H and HKU 7183/02H. In addition to the six corpora used in this study, the East African corpus is also available.
2. TACT was developed by Ian Lancashire, John Bradley, and Lidio Presutti at the University of Toronto. It is available at http://www.chass.utoronto.ca/epc/chwp/tact.html.
3. Randomization was carried out by numbering each item in each of the lists, and then using the Research Randomizer (http://www.randomizer.org/) to generate random lists of 50 numbers, with no number being selected more than once from any one list. For example, to select 50 random words from the absolute core, 50 random numbers were chosen from the numbers 1 to 11,653 inclusive.
4. *Linguistica* is available at http://www.humanities.uchicago.edu/faculty/goldsmith/index.html.
5. The list of prefixes is based on Hughes (2000: 52–3).
6. The *VocabProfile program* (also known as *Range*) can be downloaded from http://www.vuw.ac.nz/lals/publications/software.aspx.

REFERENCES

Coxhead, Averil (2000) A new academic word list. *TESOL Quarterly*, **34**(2), 213–38.
Goldsmith, John (2001) Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, **27**(2), 153–89.
Greenbaum, Sidney (1996) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.
Hughes, Geoffrey (2000) *A History of English Words*. Oxford: Blackwell.
Nation, Paul and Bauer, Laurie (1993) Word families. *International Journal of Lexicography*, **6**(4), 253–79.
Nelson, Gerald (1996) The design of the corpus. In *Comparing English Worldwide: The International Corpus of English*. Edited by Sidney Greenbaum. Oxford: Clarendon Press, pp. 27–35.
Nelson, Gerald, Wallis, Sean and Aarts, Bas (2002) *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam/Philadelphia: Benjamins.
Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey and Svartvik, Jan (1985) *A Comprehensive Grammar of the English Language*. London: Longman.
West, Michael (1953) *A General Service List of English Words*. London: Longman.

(*Received 5 October 2005.*)