languages

MDPI

Article

Indian English Evolution and Focusing Visible Through Power Laws

Vineeta Chand [1,*], Devin Kapper [2], Sumona Mondal [2], Shantanu Sur [3] and Rana D. Parshad [2]

[1] Centre for Research in Language Development throughout the Lifespan (LaDeLi), Department of Languages and Linguistics, University of Essex, Colchester CO4 3SQ, UK

[2] Department of Mathematics, Clarkson University, 8 Clarkson Ave., Potsdam, NY 13699-5815, USA; kapperdp@clarkson.edu (D.K.); smondal@clarkson.edu (S.M.); rparshad@clarkson.edu (R.D.P.)

[3] Department of Biology, Clarkson University, 8 Clarkson Ave., Potsdam, NY 13699-5815, USA; ssur@clarkson.edu

[*] Correspondence: vineeta@essex.ac.uk; Tel.: +44-01206-872101

Academic Editors: Usha Lakshmanan and Osmer Balam
Received: 19 August 2017; Accepted: 20 November 2017; Published: 24 November 2017

Abstract: New dialect emergence and focusing in language contact settings is difficult to capture and date in terms of global structural dialect stabilization. This paper explores whether diachronic power law frequency distributions can provide evidence of dialect evolution and new dialect focusing, by considering the quantitative frequency characteristics of three diachronic Indian English (IE) corpora (1970s–2008). The results demonstrate that IE consistently follows power law frequency distributions and the corpora are each best fit by Mandelbrot's Law. Diachronic changes in the constants are interpreted as evidence of lexical and syntactic collocational focusing within the process of new dialect formation. Evidence of new dialect focusing is also visible through apparent time comparison of spoken and written data. Age and gender-separated sub-corpora of the most recent corpus show minimal deviation, providing apparent time evidence for emerging IE dialect stability. From these findings, we extend the interpretation of diachronic changes in the β coefficient—as indicative of changes in the degree of synthetic/analytic structure—so that β is also sensitive to grammaticalization and changes in collocational patterns.

Keywords: language contact; power laws; World English; Indian English; diachronic; corpus linguistics; lexical diversity; dialect formation

1. Introduction

Some English varieties are unambiguously considered different 'dialects': in settings with a sustained history of institutional and home monolingual English use, e.g., the UK, North America, Australia, New Zealand, and South Africa, regional variation is attributed to distinct historic patterns of language contact [1]. However, in other non-English settings, language contact between local languages and English—which historically spread through (British) colonialism, e.g., in India and Nigeria [2], and has recently spread in conjunction with globalization, e.g., in Japan [3]—has resulted in multiple indigenized World English (WE) language contact varieties. While a growing set are now recognized as distinct English dialects, differing at all structural levels from canonical "standard" English varieties—American English and British Received Pronunciation—these WE settings and the resultant localized English patterns often have a much higher proportion of non-native (L2) speakers compared to canonical English ecologies in North America, the UK, and southern hemispheric majority English settings. WE structures and diachronic trajectories are thus influenced by processes of second language acquisition (SLA) and language contact, among other ecological factors [4]. WEs are also quantitatively and qualitatively encroaching on indigenous vernacular codes in local settings through processes of language shift, and some have reached the stage of established ideological nativization, such that the local variety surpasses the value of external norms [2].

However, the status of WEs as distinct dialects of English continues to be contentious, in large part because of their historic or continued majority L2 populations. This debate is also political, rooted in negotiations over the commodification of English(es)—and within that the (de)valuing of some forms of English, L2 and bilingual speakers of English, and non-monolingual English acquisition pathways [5,6]. The debate is further influenced by informal notions of mutual intelligibility as a criterion for separating languages: this occludes the reality that codes, while separated at extreme ends as distinct languages, often also have between those extremes a continuum of mutual intelligibility within which codes manifest as dialects. Compounding this, no specific criteria for defining or distinguishing dialects is accepted in academia. Even within WE literature, it is unclear how one could determine at what point and under what conditions a WE ecology has sufficiently developed from L2/SLA history into distinct, nativized WE dialect. Relatedly, it is an open debate how the respective role of socio-historical, ecological, demographic, and communicative factors variably influence evolutionary pathways within processes of dialect focusing and new dialect formation [2]. Indeed, language contact resulting in dialect focusing and stabilization is an ongoing process which may be (more or only) visible at a global structural level. Building on this perspective, we consider a novel lens for examining WE dialect development, focusing, and stabilization, examining the quantitative frequency characteristics of diachronic WE corpora. We next introduce these quantitative measures, then discuss how they can offer insights for capturing WE dialect development, focusing, and stabilization.

Zipf famously posited that a rank frequency distribution is a property of language, with evidence from a small written corpus (Ulysses) demonstrating that the frequency of an individual word can be derived from its ranked frequency [7]. Essentially, the distribution of words ($f(r)$) in a corpus of a particular language (where there are, say, n words), as a function of their rank (r), can be thought of as a probability mass function,

$$\sum_{r=1}^{N} f(r) = 1 \tag{1}$$

Following this, Zipf's postulate is that the distribution follows a power law given by

$$f(r) = \frac{C}{r^{\alpha}} \tag{2}$$

The α exponent is essentially the slope of the curve in a log scale, where a larger absolute α reflects a shorter tail of hapax legomena (frequency = 1) and/or low frequency words. The C and α are typically derived from the dataset/corpus in question. Intuitively (assuming α = 1), what the law says is that the second most frequent word will occur half as many times as the most frequent one, the third most frequent word will occur one third as many times as the most frequent one, and so on. His explanation, formulated as the principle of least effort, argues that non-trivial competing pressures in communication—originally presented as diversification vs. unification, later re-expressed as competition between clarity vs. ease in speaker/hearer interactions [8]—are the basis for this distributional pattern. Building on this, Mandelbrot proposed a refinement to Zipf's Law (ZL) which has come to be seen as a more general law regarding lexical frequencies in natural language [9]. According to Mandelbrot, the frequency distribution of words ($f(r)$) in a corpus of any particular language, as a function of their rank (r), is given by,

$$f(r) = \frac{C}{(\beta + r)^{\alpha}} \tag{3}$$

The β coefficient reflects deviations in the high frequency range, with a larger β reflecting a larger deviation in the curve from the α—essentially, all other things being equal, a corpus with a higher

β has comparatively less reliance on higher frequency words, while a corpus with a smaller β has an increased use of these high frequency words. Thus, in essence, Mandelbrot's law (ML) better accounts for the higher end of the ranked distribution [10]. Note in the special case $\beta = 0$, ML reduces to ZL.

Rank frequency distributions are ubiquitous in natural language, and a property of complex communicative systems: communicative optimization can both cause synchronic structural distributions and motivate diachronic statistical patterns [11–17]. Rank frequency distributions of language will have a small kernel lexicon of highly ranked words (dominated by function words and high frequency content words) and, at the opposite end, a set of words which occur only once (hapax legomena). In parallel corpora (translations of the same text), the nature of the two ends of the continuum are arguably typologically mediated: languages with more synthetic encoding strategies, i.e., with extensive inflectional systems and morphological compounding processes, will necessarily have more words which appear only once (manifesting as a longer, shallower tail of hapax legomena), while languages which rely more on analytic encoding strategies, i.e., with greater repetition of super frequent words, will have both a shorter, steeper tail of hapax legomena and a shorter, steeper slope in the high frequency range, reflecting the importance of a relatively narrow set of function words in grammatical organization [18]. These typological patterns are also visible through comparison of the model parameters across corpora: a higher absolute α exponent arguably reflects a smaller, steeper tail of hapax legomena, while a higher absolute β exponent arguably reflects a greater relative deviation from the predicted slope [18]. In comparing two data sets, Lx and Ly with originally predicted slopes αx and αy, respectively, a higher β-value for Lx suggests greater deviation in the high frequency range for Lx than for Ly. However, this difference still requires interpretation with respect to the nature of the deviation of the data from the predicted slope: e.g., for a dataset with a steeper slope in the high frequency range of the rank frequency distribution than the predicted slope, as compared to a dataset with a smaller β and a flatter, shallower slope in the high frequency range, the former can be interpreted as having comparatively higher frequencies for already high frequency words [18].

Rank frequency distributions are found in contemporary natural language corpora and Swadesh lists [19–21], comparisons across multiple languages [22–25], in both written and spoken language data [26], across all English literary texts included in Project Gutenberg [27], and historic language data that is not yet translated [28], but, importantly, are not found in random monkey-typing corpora [14,29]. Rank frequency research has expanded beyond a narrow focus on adult, monolingual, native speakers to demonstrate distinct rank frequency distributions for corpora of varying levels of L2 proficiency across users of natural language [30,31] and artificial command languages [32], L1 attritors who have lost proficiency in their L1 over their lifespan [31], different language combinations of spontaneous codeswitching [33], and in languages with varying proportions of non-native speakers [34].

One avenue of research in quantitative linguistics is to analyze language evolution through the power law constants. Diachronic analysis of the evolution of a single code uncovers specific patterns for power law constants [18,35]. Distinctive patterns have also been found in child vs. adult caregiver speech [36]: children have a lower α exponent compared to fully proficient adults. Collectively, rank frequency analysis of diachronic and comparative proficiency language data suggest that the exponent serves as an indicator of linguistic complexity [31,34–37]. Comparisons of the exponent across multiple languages which vary in their degree of synthetic to analytic structural complexity also support the relationship of the exponent to typological differences and/or to processes of typological change [33,37–39]. Based on diachronic data of the same language, rather than across different genres of synchronic data, the exponent has also been interpreted as an indicator of topical homogeneity [40]. The assimilation of SLA, L1 attrition, diachronic typologically evolving codes, and child vs. adult data collectively presents compelling evidence that comparative differences in the statistical frequency properties of different types of language data (learner vs. proficient speakers, different time periods of the same language) are robust, and can contribute to theorizing and documenting language evolution.

Yet, left unexplored in past research is how these statistical properties manifest in contemporary emerging language contact and new dialect formation, focusing, and stabilization ecologies, now understood to be far more common scenarios for language evolution and diversity than previously thought [2,4,41]. WE data offers an ideal testing ground for this. Specifically, because ZL and ML are quantitative measures of lexical diversity, with an inverse relationship, lower diversity should manifest through higher constants (C, α, β) [34]. Comparing parallel synchronic corpora, languages with a larger proportion of L2 speakers have higher constants, and less lexical diversity: adult L2 learners provoke a reduction in the number of word forms in the code diachronically [34]. In tandem, languages with more synthetic properties (more inflected forms of the same dictionary entries and less reliance on discrete function words to show grammatical properties) display lower constants, and hence greater lexical diversity [18]. However, the three constants should not simply be treated as operating in tandem: we argue that the nature of a linguistic change will affect how these constants each manifest across codes. The diachronic loss of a case system from Old to Modern English, for example, had two effects on the constants: case-marked Old English has a longer tail of hapax legomena which was lost diachronically—visible as a larger α in modern English— while modern English, relying on analytic syntactic combinations to convey the same grammatical relations, has higher frequencies for very common (typically function) words—visible as a larger β as compared to Old English [18]. While both constants increased diachronically, they were provoked by distinct grammatical features within the evolution of English.

Diachronically, we build on this to explore the frequency characteristics of new dialect formation and evolution. For WE varieties, this necessarily starts with more complicated contact between native speakers of multiple dialects of the same code (e.g., British, Scottish, and Irish English varieties) and indigenous English L2 speakers of various proficiency who also command one or more vernacular languages. Over time, WE emergence ecologically demonstrates reduced contact and influence from external English L1 speakers, a rise in the number and proportion of indigenous L2 speakers, and through this, structural changes brought about through the confluence of SLA and sustained language contact/bilingualism (compared to language shift settings) [2]. Depending on the role(s) that English plays in the local setting, the emergence of indigenous WE L1 speakers (who may or may not also command vernacular codes as bi-/multilinguals) is also possible.

As WE nativization—the development of internal norms and structures—occurs, one can predict both reductions in lexical diversity—in earlier periods process of simplification dominated by L2 speakers with input from multiple English dialects—and increases in lexical diversity—in later periods of nativization within processes of complexification by indigenous L1 speakers [42]. Pertinent to the first period posited above for WE development, L2 varieties of English and English-based pidgin/creoles commonly demonstrate reduced overt inflectional marking in zero past tense forms of regular verbs, e.g., *I talk yesterday*; a lack of inversion/auxiliaries in *wh*-questions, e.g., *Where you going?*; and a lack of number distinction in reflexives, e.g., *They saw it* [43]. Supporting this prediction, L2 population ratios also correlate with lexical diversity cross-linguistically, with lower lexical diversity in codes with larger non-native populations. Thus, in earlier periods of WE nativization, a larger α is predicted. However, the second period—of dialect focusing and stabilization—is unexplored in past literature and the focus of the current investigation. In this setting, lexical expansion could occur through processes of borrowing, new word formation, semantic shift or reallocation, and compound word formation [2]. Growth specifically in the kernel (high frequency) lexicon is also likely in later stages of L2-dominance as new collocational and syntactic patterns become established and are adopted by larger populations and lexical forms become grammaticalized and thus increase in frequency [44].

In this paper, we empirically address changes in lexical diversity within the second period of WE new dialect focusing and stabilization by comparing diachronic corpora of Indian English (IE) from 1978–2008. India boasts the largest English speaking population in the world, and, as one of the better studied English varieties, IE corpora from different time periods and L1 vs. L2 populations exist and collectively permit exploration of diachronic quantitative frequency patterns. Socially and linguistically, IE has been undergoing nativization since circa 1905 [2]. Linguistically, features of IE pertinent to exploring lexical diversity diachronically include indigenized discourse and topicalization features which have emerged through semantic reallocation (Examples (1a,b)), lexical hybrids creating compounds (1c,d), the grammaticalization of presentational focus markers (1e,f), innovative extension of transitivity patterns (1g), new verbal complement patterns (1h), different distributions of articles and plural morphemes brought about through the redistribution of the mass/count properties of some classes of nouns and quantifiers (1i–k), changes in preferences for possessive and auxiliary clitics (1l,m), distinct linguistic constraints mediating the form of grammatical case (1.n,o), variation in verb form within subjunctive constructions (1p), and lexico-grammatical innovations resulting in nativized syntactic patterns (1q,r).

1. a. Non-existential *there* [45]: *Food is there*.

   b. Invariant *isn't it* tag [45,46]: *You will eat, isn't it?*

   c. Vernacular-English compounds [47]: *tiffin-carrier*; *policewala, lathi-charged*

   d. English-English compounds [47]: *cousin-sister; cow-worship, black money, time-pass; salt giver only* focus marker [46]: *The sweets are tasty, only*.

   f. *itself* focus marker[1] [48]: *The car was purchased this year, itself.*

   g. 'new ditransitives' like *advise* occurring with two noun phrase complements in IE [49]: *I have advised him some technical changes like using both hands while stopping the ball*

   h. A markedly lower preference for NP[2] + NP complements for the ditransitive *give* in IE (~22%) vs. British English (~37%) over other complement patterns [49]

   i. Zero article [50]: *on Ø fifteenth August*; *Ø lot of X*

   j. Mass nouns pluralized [51]: *litters, furnitures, woods*

   k. Count nouns not pluralized [52]: *One of my relative Ø* . . .

   l. Lower use of uncontracted auxiliaries [53]: *will* vs. *'ll*

   m. Lower use of possessive s-genitive [51]: *I living next to my memsahib sister Ø house*.

   n. Dative as ditransitive construction *give+NP+NP* vs. prepositional construction *give+NP+PP*

   o. Genitive alternation [54]: *the party's position* vs. *the position of the party*

   p. Quantitative distinction from other Asian Englishes and British English in *was/were/would* preferences in conditional *if* clauses [55]

   q. *on* + *if* replacing *on* + *whether* in conditional *if* clauses [55]: *On if the man, indeed, was the president's security guard, he said it can be determined only after a probe*

   r. Innovative 'intrusive *as*' construction [56]: *The one who is called as Dr. Sahib* . . .

[1] These focus markers are considered calques, replicating the pattern of emphatic clitics common to many Indian languages (e.g., Hindi enclitics *hii*, and *bhii*) [46], and are examples of the grammaticalization of focus markers in IE.

[2] NP: noun phrase, PP: prepositional phrase.

This list is not exhaustive: research uncovering IE structural nativization in specific forms and constructions continues, and empirical evidence of IE structural nativization is apparent across the grammar. Structural nativization is also supported by and likely working in tandem with ideological nativization and endonormative stabilization: recent and growing evidence of such, e.g., upper class urban New Delhi IE speakers, recognizes IE as both distinct from external canonical English varieties and ideologically more valuable and appropriate in the local Indian setting [57]. Given this confluence of evidence, we are interested in what generalizations can be made from a holistic comparison of lexical rank frequency regarding IE dialect focusing, and more broadly, regarding WE nativization. This focus also speaks to a growing methodological reorientation within debates over language contact as processes of both simplification and complexification which are increasingly attending to large scale patterns [42]. While the emergence of any single new or restructured form or construction may not justify claims of nativization, holistic diachronic differences can be seen as a synthesis of evidence specifically grounded in a larger collection of such restructurings.

Within this setting, we thus ask how a quantitative exploration of power law distributions may diachronically illuminate aspects of WE new dialect focusing and contribute to understanding language contact evolution more broadly. More specifically, we explore whether there is evidence for IE lexical focusing and stabilization (e.g., towards a kernel IE lexicon reflecting increasingly cohesive syntactic and collocational patterns), which we posit will be visible through a reduction in the β diachronically. Lexical expansion, e.g., the creation of new terms, would expand the tail of hapax legomena and manifest in a smaller α: this is also likely with a diachronically growing population of native and/or fluent IE speakers who are largely speaking to other IE speakers, not to speakers of other English dialects. The goals of this paper are thus to explore, through statistical testing, which rank frequency distribution best fits this data, and to then interpret the patterns in terms of likely IE development pathways, building on previous work on diachronic change in the constants in language data [33] and the range of restructurings and innovations documented in IE.

This paper expands the scope of power law constants against diachronic language evolution to specifically consider new dialect focusing and stabilization—as such, the descriptive interpretations offered below of changes in the constants should be considered exploratory. Recognizing that language as a sophisticated system is both hierarchical and multi-faceted, we acknowledge that specific patterns within combinatorial structural levels (e.g., syntactic, semantic, pragmatic levels) will necessarily be lost in a single word frequency approach derived from transcriptions of spoken language. This approach, however, still holds value given that it is applied in tandem with and building upon linguistic analyses of other structural levels for IE and other WE varieties. We are not asking how a word frequency rank distribution is *better* than other approaches for testing new dialect stabilization, but, instead, how it may complement them, by considering how semantic, pragmatic, and morpho-syntactic innovations may also be broadly visible in distributional differences between diachronic models of lexical rank frequency within a framework that does not rely on comparison with an external canonical norm.

2. Materials and Methods

This analysis draws on two public and one private corpora (Table 1) to test our hypotheses. The Kolhapur Corpus draws on published written texts largely authored by L2 speakers: it was designed in size and genre coverage to approximately match the Brown and Lancaster-Oslo-Bergen corpora (500 texts spanning 15 categories, 2000 words per text) [58]. Next, the International Corpus of English (ICE) project has developed matching corpora covering established and new English regional settings across the globe. Each corpus includes ~1 million words spread across 500 texts from a range of registers and genres of spoken and written English. ICE-India [59] draws dominantly on L2 speakers, within which we consider the frequency characteristics of the ICE-Spoken sub-corpus separately from the ICE-Written sub-corpus. The third, and most recently collected dataset, ENDE (Elite New Delhi English) [57], draws on transcripts of spontaneous spoken life history interviews with three generations of upper class early Hindi/English bilinguals from New Delhi. ENDE includes the speech of informants (and excludes that of the interviewer), but only includes their English speech (e.g., infrequent Hindi codeswitches are excluded), and only their naturalistic speech (i.e., formal reading passage data was excluded). While the corpora have different transcription protocols for spoken data, all follow standard orthographic practices in terms of spelling conventions (e.g., *gonna* is consistently transcribed as *going to*). These corpora collectively permit a diachronic exploration of IE through (a) real-time comparison of written (comparing Kolhapur and ICE-Written) and spoken language data (comparing ICE-Spoken and ENDE); and (b) apparent time comparisons across two modalities (comparing ICE-Written and ICE-Spoken), and by speaker gender and age (separately considering the three age groups and two genders found in the ENDE corpus—discussed below and presented in Table 2).

Table 1. Details of each IE (Indian English) corpus and sub-corpus.

Corpus		Corpus Size	Year
Kolhapur Corpus	Written	1,115,139	1978
ICE-India	Written	450,857	2002
	Spoken	689,189	
ENDE	Spoken	254,530	2007-8

ICE: International Corpus of English.

Table 2. ENDE (Elite New Dehli English) age/gender stratified sub-corpus sizes. Youth are aged 18–22, Workers are aged 27–52, Retirees are aged 62–87. Because this data was collected for other purposes, the sub-corpora are not balanced in size.

	Female	Male
Youth	26,838	10,192
Workers	62,820	35,882
Retirees	65,083	53,715
Total	154,741	99,789

Within sociolinguistics, real-time analyses compare data collected at different points in time, while apparent time analyses compare synchronic snapshots of different age groups, genders, or different language modalities (i.e., spoken vs. written language, here), to extrapolate diachronic differences—and hence language change [60]. Apparent time analyses draw on the assumption that a speaker's grammar is relatively stable post-adolescence, reflecting the communicative norms from when they were age ~20, such that younger speakers use newer, emerging forms, while older speakers demonstrate older, conservative forms. Such data is less reliable evidence of diachronic change because language patterns can also change across the lifespan such that older speakers also adopt innovations [61]. However, such changes tend to emerge in forms which gain social meaning, while no lifespan-based changes in quantitative frequency patterns have been documented [61]. In gender-based apparent time comparisons, women are typically found to lead linguistic change: gender-based differences can thus provide insight into directions of change within data from the same generation [62]. Linguistic patterns more common to younger females (over older males) are interpreted as innovative, while patterns preferred by older men (over younger women) are interpreted as conservative. Similarly, linguistic innovations are understood to primarily emerge from spoken vernacular language, only later percolating into written language, such that synchronic comparisons of written and spoken data can also provide apparent time evidence of diachronic change [63]. We thus use these corpora to ask whether apparent time comparisons by gender, age, and modality support or nuance the real-time findings. Given that the ENDE corpus represents three generations of fluent bilinguals, we anticipate focusing of the β within these three age groups, and a reduced range for constants as compared to the two earlier L2 speaker-dominated corpora of IE. We also anticipate that the synchronic analysis of ICE-India Spoken and Written sub-corpora will mirror the diachronic patterns, because both are derived from L2 populations.

Apart from ZL and ML, we test two other models of rank frequency, namely the Weibull distribution, given by

$$f(r) = C\alpha\left(\frac{r^{\alpha-1}}{e^{Cr^{\alpha}\alpha}}\right) \tag{4}$$

and the Inverse-Gamma distribution given by

$$f(r) = \frac{Ce^{-b/r}}{r^{\alpha}} \tag{5}$$

Here, as usual, r denotes the rank, and α, b, and C are free parameters to be inferred from the dataset in question. The Inverse-Gamma distribution is essentially a modification of ZL.

We consider four goodness of fit measures to determine which model of rank frequency distribution best fits the data. Akaike Information Criterion (AIC) [64] is generated from a fitted model after a log-likelihood value has been obtained, according to the formula: $-2\cdot$*log-likelihood* + $k \cdot$ *npar*, where *npar* represents the number of parameters estimated in the fitted model and $k = 2$. Bayesian Information Criterion (BIC) is obtained from the same formula by letting $k = \log(n)$, where n is the number of observations used to estimate the parameters. When comparing the models fitted by maximum likelihood to the same data, smaller AIC or BIC values correspond with better fits. We also examine ΔAIC = AIC − AIC$_{\text{min}}$ to capture information lost in alternative models that deviate from the best available amongst those we considered [65]. These adjusted values are free from the scaling constants incorporated in the criterion value which is highly dependent upon sample size [65]. The Δ values are ranked such that the best available model would have a Δ value of 0, and any comparable models would be within a small tolerance. A similar analysis and interpretation is provided for ΔBIC values.

Standard error (SE) of the regression model [64] is obtained by taking the square root of the sum of the squared residual errors over all observation values and dividing the result by the degrees of freedom for the model. The last metric provided, COR(y, ŷ), represents the correlation of the log-transformed relative frequency values with their corresponding fitted values. We would like to point out that SE is not the critical indicator of fit for the power laws considered here. The AIC and BIC are more robust indicators of model fit. We note that the corpora were handled in un-lemmatized form.

Models were developed to explore various power law fits: we first tested whether rank frequency distributions could be uncovered across the corpora, and what model version provides the best fit. When considering the information loss as tracked by the ΔAIC and ΔBIC values, we see clear evidence that the Mandelbrot model provides the best fit since the Δ value is always 0. We opt to follow the rule of thumb where a model with a $\Delta > 10$ has no support for consideration [65]. We then compared the exponents across corpora and sub-corpora within the best fitting model version to consider how frequency distributions can provide insight into language evolution and new dialect focusing.

3. Results

3.1. Model Selection

A comparison of four possible power law models for the log transformed frequency data (ZL, ML, Inverse Gamma, and Weibull) demonstrates that while each one provides a good fit across the IE corpora, Mandelbrot's Law provides the best fit for every corpus (Table 3): this holds true for each of the metrics used to determine best fit: AIC, BIC, SE, and COR(y, ŷ). We base all subsequent analyses on the Mandelbrot model fit to directly compare the constants across the corpora and sub-corpora.

Table 3. Comparing model fits by corpus.[3]

	Model	AIC	ΔAIC	BIC	ΔBIC	SE	COR(y, ŷ)
Kolhapur	*Zipf*	−11,837.21	8654.40	−11,810.77	8645.58	0.215	0.9866
	Mandelbrot	**−20,491.61**	**0**	**−20,456.35**	**0**	**0.197**	**0.9888**
	Inverse Gamma	−13,249.60	7242.01	−13,214.34	7242.01	0.212	0.9870
	Weibull	−16,520.69	3970.92	−16,494.25	3962.10	0.205	0.9879
ICE-Written	*Zipf*	−10,033.62	4133.66	−10,008.53	4125.29	0.207	0.9859
	Mandelbrot	**−14,167.28**	**0**	**−14,133.82**	**0**	**0.194**	**0.9876**
	Inverse Gamma	−10,942.55	3224.73	−10,909.09	3224.73	0.204	0.9863
	Weibull	−11,246.64	2920.64	−11,221.54	2912.28	0.203	0.9866
ICE-Spoken	*Zipf*	−8108.89	5986.23	−8084.55	5978.12	0.205	0.9892
	Mandelbrot	**−14,095.12**	**0**	**−14,062.67**	**0**	**0.182**	**0.9916**
	Inverse Gamma	−9883.25	4211.87	−9850.79	4211.88	0.198	0.9900
	Weibull	−13,675.03	420.09	−13,650.69	411.98	0.183	0.9915
ENDE	*Zipf*	−4871.33	1576.10	−4849.74	1568.90	0.189	0.9909
	Mandelbrot	**−6447.43**	**0**	**−6418.64**	**0**	**0.174**	**0.9923**
	Inverse Gamma	−6020.62	426.81	−5991.83	426.81	0.178	0.9919
	Weibull	−2159.48	4287.95	−2137.89	4280.75	0.217	0.9885
ENDE: Females	*Zipf*	−3329.75	1242.85	−3309.24	1236.02	0.190	0.9906
	Mandelbrot	**−4572.60**	**0**	**−4545.26**	**0**	**0.173**	**0.9922**
	Inverse Gamma	−4304.37	268.23	−4277.03	268.23	0.177	0.9918
	Weibull	−1278.11	3294.49	−1257.61	3287.65	0.220	0.9879
ENDE: Males	*Zipf*	−3126.10	911.26	−3105.84	904.51	0.189	0.9894
	Mandelbrot	**−4037.36**	**0**	**−4010.35**	**0**	**0.176**	**0.9908**
	Inverse Gamma	−3850.87	186.49	−3823.85	186.50	0.178	0.9905
	Weibull	−1487.42	2549.94	−1467.16	2543.19	0.215	0.9869
ENDE: Youth	*Zipf*	−1297.95	869.76	−1280.23	863.85	0.190	0.9893
	Mandelbrot	**−2167.71**	**0**	**−2144.08**	**0**	**0.162**	**0.9923**
	Inverse Gamma	−1996.61	171.10	−1972.99	171.09	0.167	0.9917
	Weibull	−722.86	1444.85	−705.14	1438.94	0.212	0.9876
ENDE: Workers	*Zipf*	−2610.46	744.51	−2590.46	737.85	0.193	0.9892
	Mandelbrot	**−3354.97**	**0**	**−3328.31**	**0**	**0.181**	**0.9905**
	Inverse Gamma	−3236.01	118.96	−3209.35	118.96	0.183	0.9903
	Weibull	−928.92	2426.05	−908.93	2419.38	0.223	0.9863
ENDE: Retirees	*Zipf*	−3143.53	1164.76	−3123.20	1157.99	0.190	0.9900
	Mandelbrot	**−4308.29**	**0**	**−4281.19**	**0**	**0.173**	**0.9916**
	Inverse Gamma	−4046.35	261.94	−4019.24	261.95	0.177	0.9913
	Weibull	−1457.09	2851.20	−1436.76	2844.43	0.216	0.9876

AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; SE: standard error; COR: correlation between observed and fitted values of the log-transformed relative frequency values.

3.2. Mandelbrot's Law Constants

Given that Mandelbrot's Law provides the best fit for each corpus through each of the goodness of fit measures considered, we next explore the ML-based constants for these best fit models' quantitative statistics (Table 4) and graphically[4] (Figures 1–3), through plots of the fitted values comparing them to the log transformations of the actual frequencies.

[3] In the SE and COR(y, ŷ) columns, we round to the nearest decimal point where differences are visible.

[4] We note that the residual values are high, indicating a less than perfect fit, for the Kolhapur and ICE corpora. For all of the ENDE (sub)corpora, reasonable fit is provided for all except for a handful of the least frequent words. This is reflected in the SE values for the ENDE models, which are all lower than the models for the ICE and Kolhapur data sets.

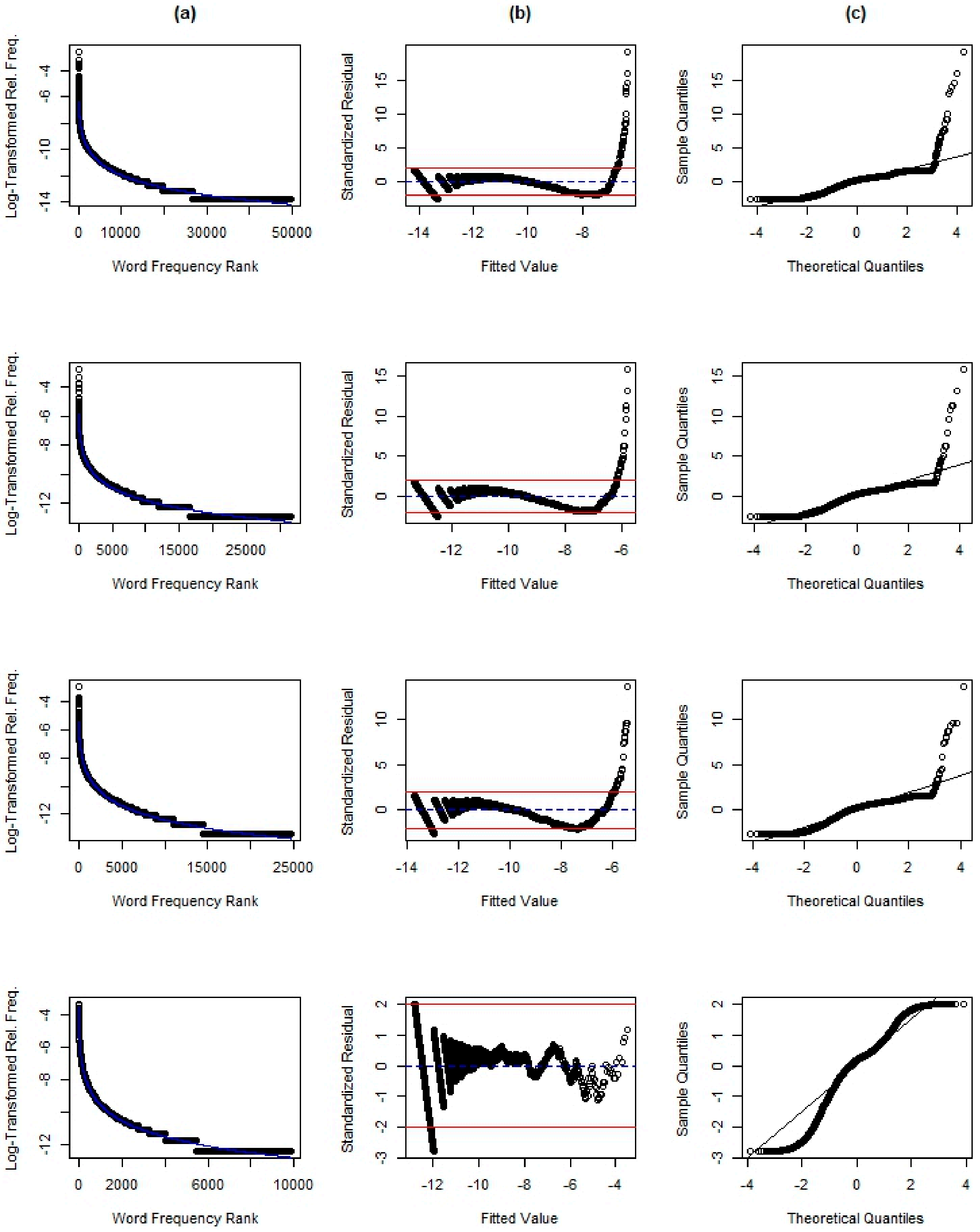

Figure 1. Analyzing Kolhapur, International Corpus of English (ICE)-Written, ICE-Spoken, and Elite New Dehli English (ENDE) corpora fit to Mandelbrot's Law. Each row corresponds to those corpora respectively. (**a**) Plot of fitted values (blue curve) relative to the observed log-transformed relative frequencies; (**b**) Standardized residuals plot with expected value of 0 indicated with a dashed blue line and extreme values at ±2 demarcated with solid red lines; (**c**) Quantile-Quantile (QQ) plot of standardized residuals with accompanying line connecting the first and third quartiles to illustrate deviations from the observed distribution. We see that the ENDE corpus has the best fit to Mandelbrot's Law, as visible by a narrower range and smaller absolute values for residuals.

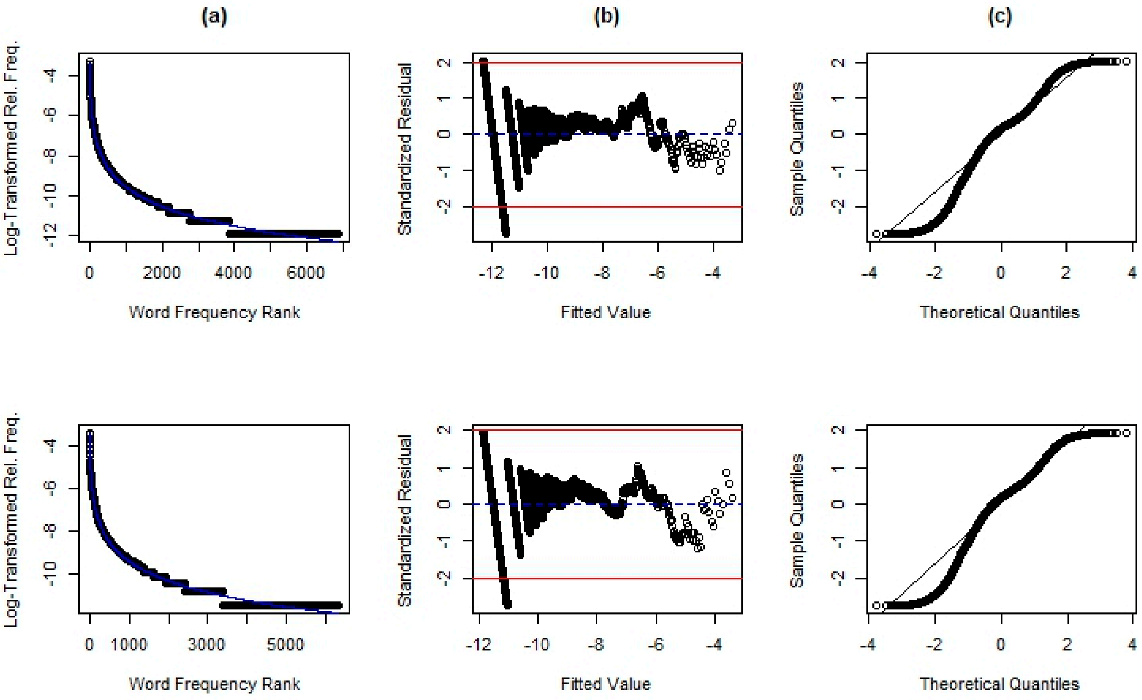

Figure 2. Analyzing ENDE Females and ENDE Males corpora fit to Mandelbrot's Law. Each row corresponds to those corpora respectively. (**a**–**c**) as described for Figure 1. For females, we do see a more random scatter of residuals (reflected in the relatively flatter slope to the right of the fitted value of −5) than for the highest ranked words, suggesting a slightly better model over the one for males.

Table 4. Comparing Mandelbrot-derived best fit models across IE corpora.

Data	Constant	α	β	COR^2(y, $\hat{y}$)
Kolhapur	2.23	1.38	176.94	0.9776
ICE-Written	1.01	1.28	91.38	0.9754
ICE-Spoken	3.44	1.48	88.33	0.9832
ENDE	1.52	1.44	14.15	0.9846
ENDE: Females	1.36	1.43	11.66	0.9844
ENDE: Males	0.85	1.34	10.35	0.9817
ENDE: Youth	1.07	1.38	10.09	0.9846
ENDE: Workers	0.86	1.35	8.59	0.9812
ENDE: Retirees	1.12	1.38	12.02	0.9833

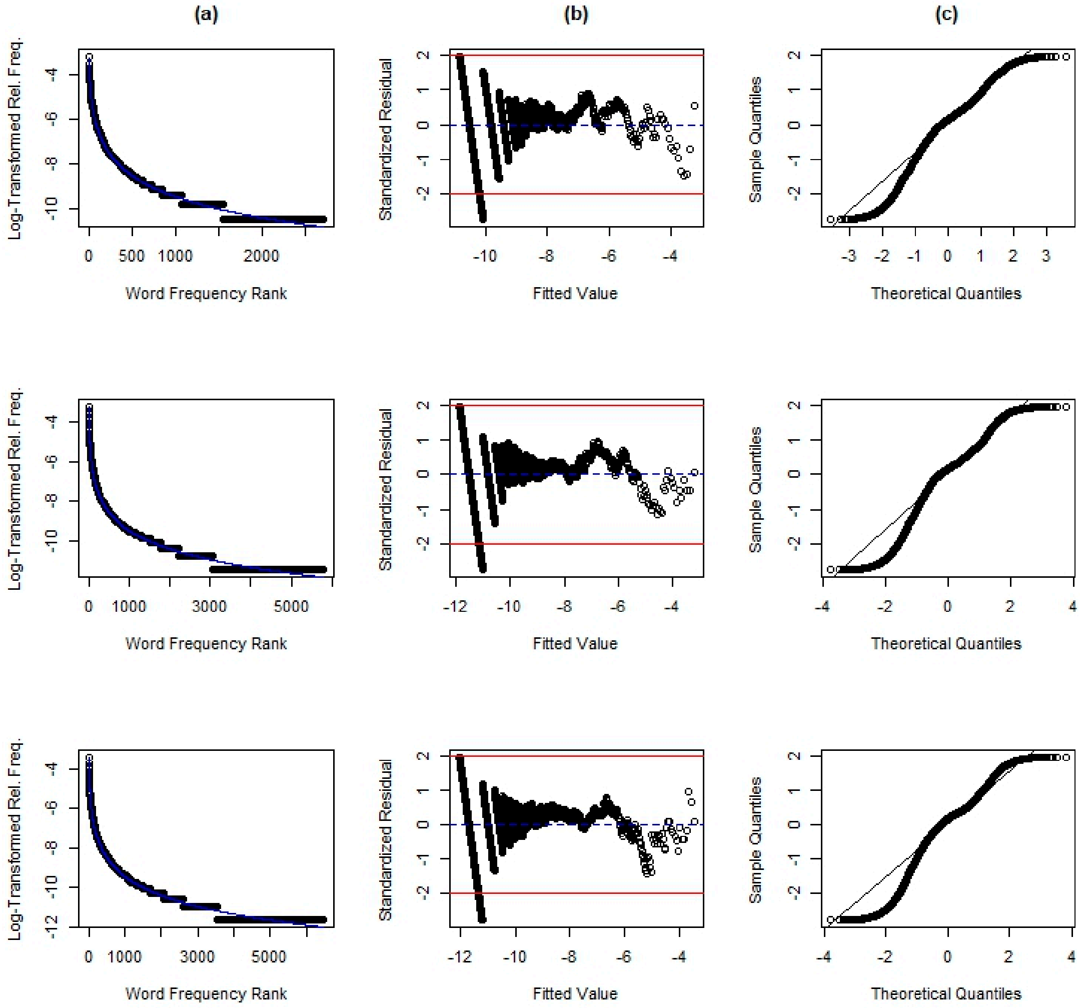

Figure 3. Analyzing ENDE Youth, ENDE Workers, and ENDE Retirees corpora fit to Mandelbrot's Law. Each row corresponds to those corpora respectively. (**a**–**c**) as described for Figure 1. The theoretical quantile range is notably smaller for youths, but that is most likely due to the much smaller subcorpus size.

4. Discussion

Each IE corpus demonstrates rank frequency distributions, uniformly best fit by Mandelbrot's law, supported by each of the goodness of fit measures we considered. Next, we discuss the quantitative characteristics of the exponents in the Mandelbrot model, α and β. We follow Bentz et al. [18] in our interpretation of the constants when possible; however, we are dealing not simply with internal linguistic change, but also with language contact-induced change.

The argument that α reflects changes in morphological marking (more pervasive inflectional systems linked to a smaller α and longer tail of hapax legomena [18]) can be extended to pidgin and other L2 learner contexts where inflections are lost (resulting in a larger α, or less lexical diversity); however, it is unclear how the α may relate to the linguistic changes common to later stages of WE dialect focusing. We hypothesize that IE dialect focusing will give rise to hapax legomena within the process of lexical expansion and may be visible in a smaller α.

In real-time, the α shrinks diachronically in both spoken (Table 4, from 1.48 to 1.44) and written (Table 4, from 1.38 to 1.28) IE modality-based comparisons, while the size of the change is larger for the written corpora. This diachronic increase in the tail of hapax legomena may be related to the increased proficiency of IE L2 speakers, and we interpret this as dialect focusing. While there is not a linear trend across the three generations of speakers in the ENDE corpus, there is minimal deviation in the α range (0.03291)—the three age groups in the most recently collected ENDE corpus are arguably quite similar and collectively are each quite small, which we interpret as evidence of dialect stabilization. Females from the ENDE corpus have a larger α than men (1.43 vs. 1.33), indicative of a smaller tail of hapax legomena. However, we note that these sub-corpora are quite small, and not well-balanced: hapax legomena show more distributional variation in smaller and topically varied corpora [66].

Comparing modalities in apparent time, the ICE- Spoken Corpus has a larger α (1.48) than the ICE-Written Corpus (1.28): this follows from the more general finding that spoken language, produced online, uses a narrower vocabulary with more high frequency words, manifesting here as a smaller tail of hapax legomena than written language, which permits time for planning and revision [63]. Overall, we interpret the apparent time modality-based difference in the α as a further indicator of dialect focusing: natural languages consistently evidence distinct differences between spoken and written language. The apparent time age-based comparison, demonstrating minimal deviation by age, and collectively a very small α, suggests recent stability and a reduction in the tail of low frequency words. While dialect evolution clearly involves lexical reduction within the establishment of a shared kernel lexicon in early stages of new dialect formation, potentially visible as a diachronic growth in the α, later stages of dialect stabilization, based on this data, involve the deployment of a larger repertoire of rare lexicon.

Diachronically, an increase in β can be interpreted as increased syntactic grammaticalization over case-based grammatical systems (a move towards a more analytic over synthetic structure [18]): while analyses of Old versus Modern English have focused on movement from case markings to periphrastic syntactic constructions, in IE, grammatical changes are evident in new, increasingly grammaticalized syntactic collocational patterns [67–69]. Dialect focusing, we hypothesize, has a more distinct kernel lexicon [70], evidenced in a smaller β diachronically. The current findings confirm this, and provide diachronic evidence for a reduction in the β coefficient in real-time in written (Kolhapur 176.9 vs. ICE-Written 91.4) and spoken (ICE-Spoken 88.3 vs. ENDE 14.1) data. We also find apparent time confirmation of this pattern across modalities in the ICE corpus, with a smaller β visible in the ICE-Spoken corpus (88.3 vs. 91.4), likely reflective of the linguistic conservatism of written language at the lexical level [71,72]. Collectively, the real and apparent time data comparing modality suggest that new dialects focus, over time, towards a shared kernel lexicon of high frequency words.

However, the apparent time comparison of ENDE age and gender sub-groups does not confirm that this trend towards grammaticalization is still in effect in the most recent data. Our interpretation rests on the minimally small range of deviance in β across the sub-corpora in concert with how small the ENDE sub-corpora β coefficients are (8.59–12.02) in comparison to the three historic corpora (Kohlapur: 176.9, ICE-Written: 91.4, ICE-Spoken: 88.3). These suggest that contemporary IE, reflected in the three apparent time age groups ENDE draws on, has stabilized after a period of focusing.

Graphically, in Figure 1, we see that the ENDE corpus has the best fit to Mandelbrot's Law, as visible by a narrower range and smaller absolute values for residuals. From Figure 2 comparing genders within the ENDE corpus, we see, for females, a more random scatter of residuals (reflected in the relatively flatter slope to the right of the fitted value of -5) for the highest ranked words, suggesting a slightly better model over the one for males. In Figure 3, comparing apparent time age groups within the ENDE corpus, the theoretical quantile range is notably smaller for youths, but that is most likely due to the much smaller sub-corpus size.

5. Conclusions

Responding to our broader research agenda, to explore WE evolution through frequency distributions, we next discuss how these real and apparent time comparisons of frequency distributions relate to new dialect formation, focusing, and stabilization within the context of language contact evolution. We found that IE does follow power law frequency distributions, based on three different corpora and considering both spoken and written language. This is a novel finding, and offers a new avenue for exploring WE dialect evolution. While innovations and restructurings in language contact settings are dominantly contrasted with canonical patterns and 'standard' codes, the current approach permits an exploration of IE diachronically without reference to an external standard. Expanding on a growing body of literature documenting discrete instances of innovation and restructuring within IE and across the cline of English contact ecologies more broadly, we examine WE nativization and focusing from a holistic lexical rank frequency perspective.

Our second interest was towards capturing and measuring diachronic dialect focusing and stabilization within language contact ecologies. In IE, we found a real-time diachronic decrease in the β coefficient for both spoken and written data and an apparent time reduction in the β from the linguistically conservative ICE-Written corpus to the innovative ICE-Spoken corpus. These changes in the β coefficient illustrate that IE has undergone lexical focusing towards a kernel lexicon and is now stable based on three generations of apparent time data. This pattern merits exploration in other WE and new dialect formation settings. Of particular interest, IE evolution (specifically, later stage focusing and stabilization) is visible over a relatively short time period (1978 to 2008), and can be captured quantitatively without unduly relying on the examination of any single or small group of linguistic features within the new dialect. These quantitative characteristics of new dialect focusing are visible in both spoken and written data, and are also visible through apparent time comparison of spoken vs. written data, while dialect stabilization is visible in the apparent time comparison of three generations of speakers in the most recent corpus.

The analysis revealed that Mandelbrot's Law provides the best fit for every corpus explored here. We see this through the AIC and BIC. We would like to point out that SE is not the critical indicator of fit for the power laws considered here.

While previous diachronic frequency research has not posited a specific explanation for changes in α, here, a real-time reduction within both spoken (Table 4, from 1.48 to 1.44) and written (Table 4, from 1.38 to 1.28) language—linguistically, longer tails of hapax legomena diachronically—arguably relates to growth in IE speakers' proficiency, one aspect of which is an increasing lexicon. This evidence further suggests that later stages of dialect stabilization involve the deployment of a larger repertoire of rare lexicon. Apparent time modality-based differences in α, with a smaller tail of hapax legomena in spoken language, are meanwhile consistent with earlier research on spoken versus written language, while this evidence extends the finding to diachronic contact situations.

Given that Mandelbrot's Law provides a better fit than Zipf's Law for these IE corpora, optimization is interpreted as an increasingly smaller β, but not specifically a β idealized as zero. We interpret this diachronic reduction of β as an indicator of internal language optimization in high frequency words. The real and apparent time evidence for a shrinking β complements previous discrete analyses encompassing a range of IE syntactic, pragmatic, and discourse-functional innovations which include the development of verb-particle collocational patterns, novel focus particles and invariant tags, and the grammaticalization of a range of lexical items [46]. These documented IE features have developed through the reshaping of individual words and structures by increasingly larger groups of fluent English speakers in India (L1 or otherwise), and we argue that their collective impact on contemporary IE is visible through frequency characteristics.

Importantly, these widespread structural changes in IE are not centered around the development or loss of a case system or specifically through a change along the cline from synthetic to analytical inflectional encoding, and instead reflect widespread individual innovations in syntactic collocational patterns, syntactic reconfigurations, and the concurrent grammaticalization of specific lexical forms to take on discourse-pragmatic functions. Based on the real and apparent time quantitative characteristics which accompany the development of these innovative and restructured IE forms, we extend previous interpretations of diachronic changes in β as reflecting a grammatical fingerprint of the inflectional state of a language [18] to argue that β also reflects the innovative functional grammaticalization of new collocational patterns. This research could profitably be extended to fit two regime models [73,74], to explicitly test whether better fits are possible, and more directly speak to distinct patterns for high versus low frequency lexicon.

This analysis also contributes to understanding how synchronic comparisons by age and gender (within ENDE), and modality (within ICE-India) can support and nuance real-time findings based on a comparison of diachronic corpora. Broadly, the ENDE sub-corpora comparisons show minimal deviation in comparison to the diachronic and modality-based comparisons. This may be because there is more genre- and register-based homogeneity within the ENDE-based oral history interviews, but it may also relate to the smaller size of the ENDE sub-corpora—future research will resolve this.

Broadly, this quantitative exploration of power law distributions in synchronic and diachronic IE data provides the first frequency characteristics for a WE. These comparisons offer evidence of language contact-based evolution and new dialect focusing and stabilization in IE which does not rely on small sets of discrete structural patterns or innovations, and also does not compare IE to external 'standard' reference points. Future analyses of additional WEs and pidgin/creoles will refine our understanding of how quantitative frequency characteristics contribute to theorizing language contact evolution.

Acknowledgments: We appreciate the thoughtful feedback of two reviewers and the editor, which have improved the paper. Support for the ENDE Corpus collection and transcription was generously provided through grants to the first author from the Wenner-Gren Foundation for Anthropological Research (#7702) and the National Science Foundation (Linguistics, #0819246).

Author Contributions: VC and RP conceived and designed the numerical experiments; DK performed the numerical experiments and analyzed the data; SM and SS contributed statistical analysis tools; VC and RP wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Trudgill, P. *Investigations in Sociohistorical Linguistics: Stories of Colonisation and Contact*; Cambridge University Press: Cambridge, UK, 2010.
2. Schneider, E.W. *Postcolonial English: Varieties around the World*; Cambridge University Press: Cambridge, UK, 2007.
3. Blommaert, J. *The Sociolinguistics of Globalization*; Cambridge University Press: Cambridge, UK, 2010.
4. Mufwene, S.S. *The Ecology of Language Evolution*; Cambridge University Press: Cambridge, UK, 2001.
5. Chand, V. [V]at is going on? Local and global ideologies about Indian English. *Lan. Soc.* **2009**, *38*, 393–419. [CrossRef]
6. Shohamy, E. Reinterpreting globalization in multilingual contexts. *Int. Multiling. Res. J.* **2007**, *1*, 127–133. [CrossRef]
7. Zipf, G.K. *Human Behavior and the Principle of Least-Effort*; Addison-Wesley: Cambridge, MA, USA, 1949.
8. Piantadosi, S.T.; Tily, H.; Gibson, E. The communicative function of ambiguity in language. *Cognition* **2011**, *122*, 280–291. [CrossRef] [PubMed]
9. Nelson, R. Statistical properties of English text produced by Korean and Chinese authors. *J. Res. Des. Stat. Linguist. Commun. Sci.* **2013**, *1*, 1–24. [CrossRef]
10. Mandelbrot, B. An informational theory of the statistical structure of language. *Commun. Theory* **1953**, *84*, 486–502.
11. Naranan, S.; Balasubrahmanyan, V.K. Models for power law relations in linguistics and information science. *J. Quant. Linguist.* **1998**, *5*, 35–61. [CrossRef]
12. Corominas-Murtra, B.; Fortuny, J.; Solé, R.V. Emergence of Zipf's law in the evolution of communication. *Phys. Rev. E* **2011**, *83*, 036115. [CrossRef] [PubMed]
13. Ferrer-i-Cancho, R. On the universality of Zipf's law for word frequencies. In *Feschrift for Altmann*; Gruyter: Berlin, Germany, 2006.
14. Ferrer-i-Cancho, R.; Elvevag, B. Random text do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* **2010**, *5*, e9411. [CrossRef] [PubMed]

15. Ferrer-i-Cancho, R.; Riordan, O.; Bollobas, B. The consequences of Zipf's law for syntax and symbolic reference. *Proc. R. Soc. B* **2005**, *272*, 561–565. [CrossRef] [PubMed]
16. Manin, D.Y. Mandelbrot's model for Zipf's Law: Can Mandelbrot's Model explain Zipf's Law for language? *J. Quant. Linguist.* **2009**, *16*, 274–285. [CrossRef]
17. Tsonis, A.A.; Schultz, C.; Tsonis, P.A. Zipf's law and the structure and evolution of languages. *Complexity* **1997**, *2*, 12–13. [CrossRef]
18. Bentz, C.; Kiela, D.; Hill, F.; Buttery, P. Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguist. Lingust. Theory* **2014**, *10*, 175–211. [CrossRef]
19. Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **2014**, *21*, 1112–1130. [CrossRef] [PubMed]
20. Tuzzi, A.; Popescu, I.-I.; Altmann, G. Zipf's laws in Italian texts. *J. Quant. Linguist.* **2009**, *16*, 354–367. [CrossRef]
21. Jayaram, B.D.; Vidya, M.N. Zipf's Law for Indian Languages. *J. Quant. Linguist.* **2008**, *15*, 293–317. [CrossRef]
22. Mehta, P.; Majumder, P. Large scale quantitative analysis of three Indo-Aryan languages. *J. Quant. Linguist.* **2016**, *23*, 109–132. [CrossRef]
23. Gelbukh, A.; Sidorov, G. Zipf and Heaps Laws' Coefficients Depend on Language. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 18–24 February 2001; Springer: Berlin/Heidelberg, Germany, 2001; pp. 332–335.
24. Bochkarev, V.V.; Lerner, E.Y.; Shevlyakova, A.V. Deviations in the Zipf and Heaps Laws in natural languages. *J. Phys. Conf. Ser.* **2014**, *490*, 012009. [CrossRef]
25. Baroni, M. Distributions in text. In *Corpus Linguistics: An International Handbook*; Lüdeling, A., Kytö, M., Eds.; Mouton de Gruyter: Berlin, Germany, 2009; Volume 2, pp. 855–873.
26. Bian, C.; Lin, R.; Zangh, X.; Ma, Q.D.Y.; Ivanov, P.C. Scaling laws and model of words organization in spoken and written language. *Europhys. Lett.* **2016**, *113*, 18002. [CrossRef]
27. Moreno-Sánchez, I.; Font-Clos, F.; Corral, A. Large-scale analysis of Zipf's law in English texts. *PLoS ONE* **2015**, 11. [CrossRef]
28. Smith, R.D. Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics* **2007**, *15*, 53–61.
29. Perline, R. Zipf's law, the central limit theorem, and the random division of the unit interval. *Phys. Rev. E* **1996**, *54*, 220. [CrossRef]
30. Laufer, B.; Nation, P. Vocabulary size and use: Lexical richness in L2 written production. *Appl. Linguist.* **1995**, *16*, 1995. [CrossRef]
31. Flores, J.M.V. Zipf's Law in L1 Attrition. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2016.
32. Ellis, S.R.; Hitchcock, R.J. The emergence of Zipf's Law: Spontaneous encoding optimization by users of a command language. *IEEE Trans. Syst. Man Cybern.* **1986**, *16*, 423–427. [CrossRef]
33. Chand, V.; Quansah, E.; Parshad, R.D.; Saha, A.; Sinha, N.; Paul, R. Zipf's law in natural spoken codeswitching: Benglish and Hinglish in India. 2017; under review.
34. Bentz, C.; Annemarie, V.; Douwe, K.; Hill, F.; Buttery, P. Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLoS ONE* **2015**, *10*, e0128254. [CrossRef] [PubMed]
35. Koplenig, A. Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes—A large-scale corpus analysis. *Corpus Linguist. Linguist. Theory* **2015**. [CrossRef]
36. Baixeries, J.; Elvevag, B.; Ferrer-i-Cancho, R. The evolution of the exponent of Zipf's law in Language Ontogeny. *PLoS ONE* **2013**, *8*, 1–14. [CrossRef] [PubMed]
37. Bentz, C. Zipf's law and the grammar of languages: Synthetic and analytic encoding strategies across languages of the world. Available online: http://www.christianbentz.de/Posters/ChrisBentz_Poster.pdf (accessed on 23 November 2017).
38. Corral, A.; Boleda, G.; Ferrer-i-Cancho, R. Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS ONE* **2015**, *10*, e0129031. [CrossRef] [PubMed]
39. Zanette, D.; Montemurro, M. Dynamics of text generation with realistic Zipf's distribution. *J. Quant. Linguist.* **2005**, *12*, 29–40. [CrossRef]
40. Williams, J.R.; Bagrow, J.P.; Reagan, A.J.; Alajajian, S.E.; Danforth, C.M.; Dodds, P.S. Zipf's law is a consequence of coherent language production. *arXiv*, 2016.
41. Blommaert, J.; Collins, J.; Slembrouck, S. Spaces of multilingualism. *Lang. Commun.* **2005**, *25*, 197–216. [CrossRef]

42. Kortmann, B.; Szmrecsanyi, B. World Englishes between Simplification and Complexification. In *World Englishes Problems, Properties and Prospects: Selected Papers from the 13th IAWE Conference*; John Benjamins Publishing: Amsterdam, The Netherlands, 2009; p. 265.
43. Kortmann, B.; Schneider, E.W. *A Handbook of Varieties of English: A Multimedia Reference Tool*; Mouton de Gruyter: Berlin, Germany, 2004.
44. Hilbert, M. Interrogative inversion in non-standard varieties of English. *Lang. Contact Contact Lang.* **2008**, *7*, 261.
45. Nihalani, P.; Tongue, R.K.; Hosali, P. *Indian and British English: A Handbook of Usage and Pronunciation*; Oxford University Press: Delhi, India, 1979.
46. Lange, C. *The Syntax of Spoken Indian English*; John Benjamins: Amsterdam, The Netherlands, 2012.
47. Kachru, B.B. The Indianness in Indian English. *Word* **1965**, *21*, 391–410. [CrossRef]
48. Sridhar, K.K. The pragmatics of South Asian English. In *South Asian English: Structure, Use and Users*; Baumgardner, R.J., Ed.; University of Illinois Press: Urbana, IL, USA, 1996; pp. 141–157.
49. Hoffmann, S.; Mukherjee, J. Ditransitive verbs in Indian English and British English: A corpus-linguistic study. *AAA Arbeiten aus Anglistik und Amerikanistik* **2007**, *32*, 5–24.
50. Kachru, B.B. South Asian English. In *English as a World Language*; Bailey, R.W., Gorlach, M., Eds.; University of Michigan Press: Ann Arbor, MI, USA, 1982; pp. 353–383.
51. Hosali, P.; Atchison, J. Butler English: A minimal pidgin? *J. Pidgin Creole Lang.* **1986**, *1*, 51–79. [CrossRef]
52. Chelliah, S.L. Constructs of Indian English in language 'guidebooks'. *World Engl.* **2001**, *20*, 161–177. [CrossRef]
53. Sharma, D. The pluperfect in native and non-native English: A comparative corpus study. *Lang. Var. Chang.* **2001**, *13*, 343–373. [CrossRef]
54. Heller, B.; Bernaisch, T.; Gries, S.T. Empirical perspectives on two potential epicenters: The genitive alternation in Asian Englishes. *ICAME J.* **2017**, *41*, 111–144. [CrossRef]
55. Hundt, M.; Hoffmann, S.; Mukherjee, J. The hypothetical subjunctive in South Asian Englishes: Local developments in the use of a global construction. *Engl. World-Wide* **2012**, *33*, 147–164. [CrossRef]
56. Lange, C. The 'intrusive as'-construction in South Asian varieties of English. *World Engl.* **2016**, *35*, 133–146. [CrossRef]
57. Chand, V. Who owns English? Political, Social and Linguistic Dimensions of Urban Indian English Language Practices. Ph.D. Dissertation, University of California, Davis, CA, USA, 2009.
58. Shastri, S.V.; Patilkulkarni, C.T.; Shastri, G.S. *Manual of Information to Accompany the Kolhapur Corpus of Indian English, for Use with Digital Computers*; Department of English, Shivaji University: Kolhapur, India, 1986.
59. Shastri, S. Overview of the Indian component of the international corpus of English (ICE-India). In *Attachment of ICE-Indian Corpus*; Shivaji University, Freie Universitat Berlin: Kolhapur, India, 2002.
60. Bailey, G. Real and apparent time. In *The Handbook of Language Variation and Change*; Chambers, J.K., Trudgill, P., Schilling-Estes, N., Eds.; Blackwell: Oxford, UK, 2002; pp. 312–332.
61. Wagner, S.E. Age grading in sociolinguistic theory. *Lang. Linguist. Compass* **2012**, *6*, 371–382. [CrossRef]
62. Labov, W. *Principles of Linguistic Change: Social Factors*; Basil Blackwell: Oxford, UK, 2001; Volume 2.
63. Chafe, W.; Tannen, D. The relation between written and spoken language. *Annu. Rev. Anthropol.* **1987**, *16*, 383–407. [CrossRef]
64. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Statistical Models*; Irwin Chicago: Chicago, IL, USA, 1996; Volume 4.
65. Burnham, K.P.; Anderson, D.R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]
66. Biber, D. Representativeness in corpus design. *Lit. Linguist. Comput.* **1993**, *8*, 243–257. [CrossRef]
67. Balasubramanian, C. *Register Variation in Indian English*; John Benjamins: Amsterdam, The Netherlands, 2009.
68. Schneider, E.W. How to trace structural nativization: Particle verbs in world Englishes. *World Engl.* **2004**, *23*, 227–249. [CrossRef]
69. Sedlatschek, A. *Contemporary Indian English: Variation and Change*; John Benjamins: Amsterdam, The Netherlands, 2009.
70. Ferrer i Cancho, R.; Solé, R.V. The small world of human language. *Proc. R. Soc. Lond. B Biol. Sci.* **2001**, *268*, 2261–2265. [CrossRef] [PubMed]

71. Labov, W. *The Social Stratification of English in New York City*; Center for Applied Linguistics: Washington, DC, USA, 1966.
72. Labov, W. *Principles of Linguistic Change: Internal Factors*; Blackwell: Oxford, UK, 1994.
73. Ferrer-i-Cancho, R.; Solé, R.V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.* **2001**, *8*, 165–173. [CrossRef]
74. Li, W.; Miramontes, P.; Cocho, G. Fitting ranked linguistic data with two-parameter functions. *Entropy* **2010**, *12*, 1743–1764. [CrossRef]