BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Frequencies of Phonemes in the General British and General American Accents of English: Literature Review and New Estimates

Andrei Gonzales Iturri [1] **,** ***Greg Brooks*** [2] **,** ***Bernard Baycroft*** [3] **,** ***Gill Cochrane*** [4*]

[1] *Department of Phonetics, Spell Bee International, La Paz 00591, Bolivia*

[2] *Formerly School of Education, University of Sheffield, S10 2TN Sheffield, UK*

[3] *Department of Spanish and Portuguese (Emeritus Professor), Stanford University, Stanford, CA 94305, USA*

[4] *Training Department, Dyslexia Action, TW18 4AX Staines, UK*

ABSTRACT

This article analyses the frequency of phonemes in the General British (GB) and General American (GA) accents of English in their contemporary forms in adult usage. Part One presents a critical analysis of previous data on British and American accents, much of which data are phonetically deficient. We reveal the absence of any phonetically rigorous analysis of the GA accent. Part Three presents a thorough new analysis of the frequency of phonemes in GB and GA based on the Corpus of Contemporary American English (COCA): a corpus containing over 1 billion words. The first 20,200 most frequently occurring words in COCA were transformed into iteratively justified International Phonetic Alphabet (IPA) transcriptions in both accents. Results are shown for both lexical frequencies and text frequencies and for analyses based on the first 1000, 2000, 3000 and 4000 most frequent words in COCA. The analyses of both frequency types in both accents provide a much-needed source of up-to-date information about English language usage that fills critical accent inventory lacunae. The data will enable literacy scheme designers and practitioners involved in teaching English as an additional language to optimise the structure of their programmes to ensure learners access the most productive phonemes at the earliest juncture. Those working within the field of assistive technology and other emerging speech-communication technologies will also benefit from access to this contemporary catalogue of the GA and GB accents.

Keywords: Phoneme Text Frequency; Phoneme Lexical Frequency; GB Accent; GA Accent; Phonics Scheme Design

*CORRESPONDING AUTHOR:

Gill Cochrane, Training Department, Dyslexia Action, TW18 4AX Staines, UK; Email: gcochrane@dyslexiaaction.org.uk

ARTICLE INFO

Received: 2 July 2025 | Revised: 28 September 2025 | Accepted: 15 October 2025 | Published Online: 12 November 2025
DOI: https://doi.org/10.30564/fls.v7i12.10858

CITATION

Gonzalez Iturri, A, Brooks, G., Baycroft, B., et al., 2025. Frequencies of Phonemes in the General British and General American Accents of English: Literature Review and New Estimates. Forum for Linguistic Studies. 7(12): 960–989. DOI: https://doi.org/10.30564/fls.v7i12.10858

1. Assumptions and Literature Review

1.1. Aim

Our intention in this article is to make available reliable new estimates of the frequencies of the 44 phonemes of the General British (GB) accent of English and of the overlapping but not fully coterminous 44 phonemes of the General American (GA) accent of English in adult usage. We focus on GB and GA because phoneticians have extensively studied these accents, and because they are perceived as having status in the United Kingdom and the United States, respectively. 'General British' is now phoneticians' and linguists' preferred term for the British accent previously known as 'Received Pronunciation' or RP.

Such frequencies are intended to be of interest and use to phoneticians and linguists more generally, and to provide data on possible shifts in the two accents relative to previous analyses. Also, new rank orders for the frequencies of the phonemes in the two accents may provide useful information for those devising phonics materials for the teaching of initial literacy and those involved in speech-communication technology.

Anticipating our conclusions somewhat, we can say that the most recent reliable analysis of the GB accent appears to be Knowles [1]; and, bizarrely, that there appears to be no reliable analysis ever of the GA accent as such on a scientific 44-phoneme basis. Even the earliest attempts to identify the set of General American (GA) phonemes did not clearly distinguish them from those of the General British (GB) accent [2].

Such frequencies are also needed as the basis for calculations of the frequencies of phoneme-grapheme and grapheme-phoneme correspondences relative to the British and American spelling systems of English, respectively. We intend the data presented here to lead to re-calculations of those frequencies, to replace those of Carney [3] and P. F. D. Gontijo, I. Gontijo and R. Shillcock [4] for GB, and as the basis for the first such calculations for GA. From there, the data would also be incorporated in the second edition of Brooks [5] for GB and in Brooks and Baycroft (in preparation) for GA, respectively.

New estimates of phoneme frequencies were thought to be needed because existing ones were out of date and/or based on limited language samples. The most recent count of phonemes in GB appears to be Knowles [1]. As recently as the 8th edition (2014) of *Gimson's Pronunciation of English* [6], the editor, Alan Cruttenden, used data from Knowles [1]; if there had been more recent data, it seems likely he would have used them. As we will show later in this article, we have not been able to locate phoneme data for the GA accent as such — none of the available sources cover the full tally of 44 phonemes specified within this article.

1.2. Phonetic Preliminaries

International Phonetic Alphabet (IPA) symbols are used in this article; they are explained in **Tables 1, 2** and **3**. Phonemes are enclosed in double forward slashes, / /, as has been standard practice in phonetics for more than a century, and graphemes are enclosed in paired angle brackets, < >; in mathematics, these are better known as the signs for "is less than" and "is greater than"; no implication of those meanings is intended here. (This convention dates back at least to P. R. Hanna, J. S. Hanna, Hodges and Rudorf [7] in 1966). 'Vowel' and 'consonant' refer to phonemes, not letters. Vowels comprise both pure vowels and diphthongs. Long and short vowels are also defined phonetically, and not by the traditional use of those terms in literacy teaching. For the purposes of this article, and based on the relevant phonetic/linguistic literature, the number of phonemes in relevant categories in the two accents are those shown in **Table 1**.

The IPA symbols for the 24 consonant phonemes common to the two accents are those shown in **Table 2**.

The IPA symbols for the (partly shared, partly different) vowel phonemes in the two accents (20 in each) are those shown in **Table 3**.

Table 1. Categories of phonemes in the General British and General American accents.

			GB	GA
Consonants	Voiced		15	15
	Voiceless		9	9
		Subtotal	**24**	**24**

Table 1. *Cont.*

Pure vowels	Short		7	6
	Long		5	3
Diphthongs			8	5
Retroflex (‘r-coloured’) vowels			0	6
		Subtotal	**20**	**20**
		TOTAL	**44**	**44**

Table 2. The International Phonetic Alphabet symbols for the 24 consonant phonemes of the General British and General American accents of English.

Phoneme	Where Phoneme Occurs within Exemplar Word	Exemplar Word	Phonetic Transcription
/b/	as in the first sound of	*By*	/baɪ/
/d/	as in the first sound of	*Dye*	/daɪ/
/g/	as in the first sound of	*Goo*	/guː/
/m/	as in the first sound of	*My*	/maɪ/
/n/	as in the first sound of	*Nigh*	/naɪ/
/p/	as in the first sound of	*Pie*	/paɪ/
/t/	as in the first sound of	*Tie*	/taɪ/
/r/	as in the first sound of	*Rye*	/raɪ/
/k/	as in the first sound of	*Coo*	/kuː/
/ʧ/	as in the first sound of	*Chew*	/ʧuː/
/f/	as in the first sound of	*Few*	/fjuː/
/ʤ/	as in the first sound of	*Jaw*	/ʤɔː/
/l/	as in the first sound of	*Law*	/lɔː/
/s/	as in the first sound of	*Sue*	/suː/
/v/	as in the first sound of	*View*	/vjuː/
/z/	as in the first sound of	*Zoo*	/zuː/
/h/	as in the first sound of	*Who*	/huː/
/ŋ/	as in the **last** sound of	*Ring*	/rɪŋ/
/ʃ/	as in the **third** sound of	*Fission*	/ˈfɪʃ. ən/
/ʒ/	as in the **third** sound of	*Vision*	/ˈvɪʒ.ən/
/θ/	as in the first sound of	*Thigh*	/θaɪ/
/ð/	as in the first sound of	*Thy*	/ðaɪ/
/w/	as in the first sound of	*Well*	/wel/
/j/	as in the first sound of	*yell, union*	/jel, ˈjuːnj.ən/

Table 3. The International Phonetic Alphabet symbols for the 20 vowel phonemes of the GB accent of English and for the 20 vowel phonemes of the GA accent of English.

Symbols		Examples		Transcriptions	
GB	**GA**			**GB**	**GA**
Short pure vowels: /æ ɛ ɪ ʌ ʊ ə/, plus GB /ɒ/					
/æ/	/æ/	as in the first sound of	*Ant*	/ænt/	/ænt/
/e/	/ɛ/	as in the first sound of	*End*	/end/	/ɛnd/
/ɪ/	/ɪ/	as in the first sound of	*Ink*	/ɪŋk/	/ɪŋk/
/ɒ/		as in the first sound of	*Odd*	/ɒd/	
/ʌ/	/ʌ/	as in the first sound of	*Up*	/ʌp/	/ʌp/
/ʊ/	/ʊ/	as in the second sound of	*Pull*	/pʊl/	/pʊl/
/ə/ (schwa)	/ə/ (schwa)	as in the first sound of and the last sound of	*about* *drama*	/ə’baʊt/ /ˈdrɑːmə/	/ə’baʊt/ /ˈdrɑːmə/
Long pure vowels: /ɑː iː uː/, plus GB /ɜː ɔː/					
/ɑː/	/ɑː/	as in the whole sound of	*ah!*	/ɑː/	/ɑː/
/iː/	/iː/	as in the first sound of	*Eel*	/iːl/	/iːl/
/uː/	/uː/	as in the first sound of	*Ooze*	/uːz/	/uːz/

Table 3. *Cont.*

Symbols		Examples		Transcriptions	
GB	**GA**			**GB**	**GA**
Long pure vowels: /ɑː iː uː/, plus GB /ɜː ɔː/					
/ɜː/		as in the last sound of	*Milieu*	/miːlˈjɜː/	
/ɔː/		as in the whole sound of	*Awe*	/ɔː/	
Retroflex ('r-coloured') vowels (GA only)					
	/ɑɚ/	as in the whole sound of	*Are*		/ɑɚ/
	/eɚ/	as in the whole sound of	*Air*		/eɚ/
	/ɪɚ/	as in the whole sound of	*Ear*		/ɪɚ/
	/oɚ/	as in the whole sound of	*Ore*		/oɚ/
	/ʊɚ/	as in the second sound of	*Tour*		/tʊɚ/
	/ɝ/	as in the second sound of	*bird*		/bɝd/
		and the last sound of	*oyster*		/ˈoɪstɝ/
Other (non-retroflex) diphthongs: /eɪ aɪ əʊ (oʊ) aʊ ɔɪ (oɪ)/, plus GB /eə ɪə ʊə/					
/eɪ/	/eɪ/	as in the first sound of	*Aim*	/eɪm/	/eɪm/
/aɪ/	/aɪ/	as in the first sound of	*Ice*	/aɪs/	/aɪs/
/əʊ/	/oʊ/	as in the first sound of	*Oath*	/əʊθ/	/oʊθ/
/aʊ/	/aʊ/	as in the first sound of	*Ouch*	/aʊʧ/	/aʊʧ/
/ɔɪ/	/oɪ/	as in the first sound of	*Oil*	/ɔɪ/	/oɪ/
/eə/		as in the whole sound of	*Air*	/eə/	
/ɪə/		as in the whole sound of	*Ear*	/ɪə/	
/ʊə/		as in the second sound of	*Rural*	/ˈrʊərəl/	

Our source for the 44 phonemes of GB is Gimson's *Pronunciation of English*. All 8 editions (1962–2014) list the same set of phonemes with the same symbols throughout, for example, Gimson[8], and we have seen no reason to diverge from that analysis, with one significant exception. At the end of Part Three, for reasons given there, we separate out the two-phoneme sequence /juː/, accord it quasi-phoneme status, and provide an analysis based on the resulting 45-item set. The same 44-phoneme set was also used in all three of the previous text frequency analysis cited below: Fry[9], Denes[10,11] and Knowles[1]. Our source for the 44 phonemes of GA is *Merriam-Webster's Advanced Learner's English Dictionary*[12], specifically the table shown on pages 22a & 1994. The third author contacted Merriam-Webster dictionaries to enquire about an academic source for this phoneme set. A member of their editorial team (personal communication, 9/1/25) provided the following information: "Our system for displaying the pronunciation of words is a proprietary system that was developed by our in-house linguists. The most thorough discussion of this system can be found in *Merriam-Webster's Collegiate Dictionary, Eleventh Edition*[13], in the front-matter section titled 'A Guide to Pronunciation' (p. 33a)."

While much of the information in **Tables 1–3** seems to us uncontroversial, some points require comment. First, consonants are practically equivalent between accents, despite GA /t/ (in its flapped realisation [ɾ]) often seeming close to [d], we have treated it as equivalent to GB /t/. We treat the alleged syllabic consonants /l̩, m̩, n̩/ as sequences of schwa /ə/ plus the plain consonant /əl əm ən/, respectively. Similarly, for vowels we treat GB /e, əʊ, ɔɪ/ as allophonic with respect to GA /ɛ, oʊ, oɪ/ respectively, and ignore the difference between the accents in the quality of /æ/. We retain /ʊə, ʊɚ/ because the disappearance of the former in GB is not yet total as an R-linked vowel, and the latter seems not to be declining in GA. The (Californian) third author felt he had been thrown a curve ball when the (British) second author pointed out that GA has no short vowel phoneme corresponding to the letter <o>, despite the intuitions and deep-rooted teaching tradition of teachers in the US. He accepted that, where GB has /ɒ/, GA mainly has /ɑː/ — for justification, see Gimson[6,8], (under /ɒ/); Johnson and Ladefoged[14] (pp. 41–43 and Table 2.2). We consider GA stressed /ɝ/ and unstressed /ɚ/ as allophones like in the word *burger* being /ɜː, ə/ respectively in GB, and see no reason to differentiate them (as, for example, do Hanna et al.[7]; Mines, Hanson & Shoup[15]). Following Brooks[5] (pp. 101–103, 106–108) we recognise

and transcribe multiple instances of /w, j/ which have no representation in spelling but are present in speech as automatic glides between vowels which would otherwise be in hiatus. Examples include *doing* /ˈduːw.ɪŋ/, *laity* /ˈleɪ.jɪ.tiː/.

The two most noticeable differences between the GB and GA accents are in syncopation and rhotacism. By syncopation is meant the tendency for words of three or more syllables to lose a syllable, which is a strong feature of GB, but not of GA – for example, *incendiary* has four syllables in GB /ɪnˈsen.ʤə.riː/, five in GA /ɪnˈsen.diː.jeɚ.riː/. Rhotacism is present in GA, absent in GB (see, for example, /ə, eɚ/, corresponding to letter <a> in *incendiary*), such that the five retroflex diphthongs and one simple retroflex vowel of GA shown in **Table 3** have some parallels, but not close equivalents, in GB. Where in GA a retroflex vowel is followed (in spelling) by <r> and a vowel letter, we make the intervening /r/ phoneme explicit; this applies particularly to words ending in <-ary, -ery, -ory>.

There are no apparent differences in the following vowels between accents: /iː, ɪ, ʌ, uː, ʊ, eɪ, aɪ, aʊ/. The main differences in the analysis of vowel frequencies can be identified as follows: GB /ɑː/ corresponds to GA /ɑː, ɑɚ/, with a few instances of GB /ɑː/ corresponding to GA /æ/. GB /ɔː/ corresponds to GA /ɑː, oɚ/ and GB /ɒ/ corresponds to GA /ɑː/, as mentioned previously. Additionally, GB /eə, ɪə, ʊə/ correspond to GA /eɚ, ɪɚ, ʊɚ/.

1.2.1. Outline Summary of Results of Literature Search

The authors based their search of the relevant literature on the 44-phoneme lists for GB and GA. (In brief, for fuller details see below), we identified just 7 previous analyses whose data we considered rigorous enough to analyse and tabulate: three for a US accent with dates ranging from 1874 to 1966, and four for the GB accent with dates ranging from 1935 to 1987. Our first conclusion, therefore, is that new data are most certainly needed. We also found that none of the three US analyses mentioned below deals, strictly speaking, with the GA accent in the sense of being based on the 44-phoneme set which we have identified for that accent.

1.2.2. Cross Checking the Absence of Data on the General American Accent

The absence of such data was so puzzling that we undertook three further convergent searches. Firstly, the second author contacted two noted authorities in the field: Professor P. David Pearson in California, and Professor of Phonetics Jane Setter at the University of Reading in the UK. He asked them whether they knew of relevant and recent analyses of the phonemes of the general American accent; both said they knew of no such information.

Secondly, the second author requested a librarian at the University of Sheffield, UK, to conduct a systematic search of the scientific literature. The search strategy, which was eventually agreed upon between the second author and the university librarian, is reproduced as **Appendix A** to this article. The databases searched were SCOPUS and Web of Science. The second author trawled through the results and found nothing of relevance to this article.

Finally, in 2024, we contacted Prof Keith Johnson, co-author with the late Peter Ladefoged, of *A Course in Phonetics*, which has been a standard text in the United States for many years (various editions 1975–2015). He kindly shared some unpublished data of his own (personal communication, 12/9/24), which we have decided not to use, partly because we have otherwise used only published data, and partly because his phoneme inventory does not contain the full set of six retroflex vowels. We therefore proceeded on the assumption that only Whitney[16,17] and Dewey[18] provided reliable text frequency data on US accents, and that only Hanna et al.[7] provided reliable lexical data on a US accent. If the wider linguistic and phonetic community has relevant data, the present authors would be glad to hear from them. Also, why so many American phoneticians have used a set of phonemes with significantly fewer than 44 members (including those who designed the *Pronouncing Dictionary of American English,* Kenyon and Knott[19] would warrant a detailed history and analysis, but that task is not undertaken here.

This has been the case for many years; several books have taught the General American (GA) accent using an unclearly distinguished set of phonemes. Even today, for example, Carley and Mees[20] (pp. 125−135) describe a set that merges /ʌ/ and /ə/ and mixes GB /ɔː/ with an unusual phoneme like /o/ (Cardinal Vowel No. 7 — which is not present in either GB or GA). A major issue is the omission of rhotic vowels, where GB has /ɑː+r/, /ɜː+r/, /ɔː+r/, /eə, ɪə, ʊə/, they had given the GA /ɑr/, /ər/, /or/, /ɛr, ir, ur/ without featuring them with their respective proper rhotic symbols. The

authors chose to exclude these variations in their analysis, as shown previously in **Table 3**.

1.3. Calculating Phoneme Frequencies from Written Vs Spoken Sources, or Both

Some authors of phoneme frequency calculations have used written sources transcribed phonetically as their material (e.g., Fry[9]). This approach arose because written texts were much more available and accessible than recordings of speech, but it also requires the assumption that phonetically transcribed written materials can legitimately represent the spoken language. Other authors maintain that the natural rate of occurrence of phonemes can be validly investigated only in spoken discourse (e.g., Denes[10,11]).

However, it appears that this divergence is unwarranted. Fowler[21] concluded that "The distributional arrangements of phonemes are a part of the structure of the language which is not significantly disturbed by any individual differences in either author or subjec" (p. 47). Wang and Crawford's[22] analysis also "indicate[d] relatively high comparability between conversational and non-conversational materials" (p. 137), even though there were wide differences in style between very formal literary prose and colloquial telephone conversations. Gerber and Vertin[23] suggested a reason for the lack of difference: "The statistical constraints upon a particular language are so severe that variations in time, place or form are of little consequence" (p. 140). A similar conclusion was reached by Mines et al.[15] (p. 223), for example. In this article, we therefore consider all relevant frequency calculations, whether derived from purely written sources or from purely spoken sources or (as is the case in many of the publications reviewed) from a blend of the two.

In the rest of this article, we consider (1) text frequency data on US accents; (2) text frequency data on the GB accent; (3) lexical frequencies across both; and (4) our own new data, which cover both lexical and text frequencies.

1.4. Text Frequencies and Lexical Frequencies

These two types of phoneme frequency were first clearly distinguished by Trubetzkoy[24], (pp. 256, 257), though the distinction was implicit in less clear terminology in Trnka[25]. Both types of frequency feature in the literature, and can be defined as follows:

- Text frequencies are the frequencies with which phonemes occur when all instances of them in a language corpus are counted. Since some words will occur many times, text frequencies take into account the frequencies of the words in which the phonemes occur. Text frequencies are therefore based on word tokens.
- Lexical frequencies are the frequencies with which phonemes occur when all and only the instances of them in the phonetic transcriptions in a word list (e.g., a dictionary) are counted. Since words occur only once each in a word list or dictionary, lexical frequencies take no account of the frequencies of the words in the general language. Lexical frequencies are therefore based on word types.

Usually, the two frequencies are similar, but where a phoneme occurs in only a few words, but those words are very common, the text frequency will be high and the lexical frequency low, and vice versa, where a correspondence occurs in many words but those words are rare. The clearest example is the definite article *the*, which is the most frequent word in English by a wide margin. For lexical frequencies, it counts once; for text frequencies, it will count many thousands or even millions of times, so its constituent phonemes, /ð, ə/, also count as often. Another example is the case of /ʊə/ having higher frequencies in both accents in lexical frequencies and lower in text frequencies. Wang and Crawford[22] demonstrated convincingly that the two types of analysis can produce significantly different frequencies and correlate poorly — at least for consonants — they did not cover vowels.

That the correlation is weak for both consonants and vowels can be deduced from Berndt, Reggia & Mitchum[26]. In their Appendix C (p. 9), they list the rank order of the lexical phoneme frequencies for the GA accent produced by Hanna et al.[7] from written sources, and alongside that, the rank order for the text phoneme frequencies for the GB accent produced by Denes[10,11] from transcribed speech. It might be considered odd to draw a conclusion from a correlation between written and spoken phoneme frequency rank orders, but inspection of the wide discrepancies in ranks for individual phonemes should dispel this misgiving. For ex-

ample, /r/ has rank 1 in Hanna et al.[7], rank 12 in Denes; /æ/ has ranks 10 and 24, respectively. The largest disparity is for /ð/: 38 vs 9 — as Berndt et al.[26] point out (p. 3), this is obviously because, although /ð/ occurs in only 149 words in Hanna et al.'s[7] corpus of 17,310 words, some of those are of very high frequency in the language, especially *the*. Consequently, Berndt et al.[26] reported (p. 4) a Spearman rank correlation between the two rank orders of only 0.65.

In this article, we deal with both text and lexical frequencies, in that order, and mainly separately because none of the previous authorities quoted reported both, but in reporting our own results, we show both and compare them.

1.5. Phoneme Text Frequencies in US Accents

Following reference trails uncovered about a dozen attempts to calculate phoneme text frequencies in US accents. However, most had to be discarded. Atkins[27] was said by Hayden[28] to have "made a count of the speech sounds in Thorndike's word list" (p. 218, Thorndike[29]), but inspection of Atkins's article shows her 'speech sounds' were a mixture of a few phonemes and mainly two-letter items, therefore unusable here. A study based on the Rhode Island accent was not generalisable enough (Agard's data in Carroll[30] — for a brief description of this and further reasons for discarding it, see Wang[31]). Wang and Crawford[22] and Gerber and Vertin[23] analysed various previous authors' data but presented no new findings. Two early attempts to use spoken data (French, Carter & Koenig[32]; c.f., French and Koenig[33]; Voelker[34]) were, as pointed out by Hayden[28], based on inadequate samples and/or a mixture of accents. Moreover, French et al.[32] screened out not only names, hesitation noises, profanities, etc., but also "the articles 'the', 'a' and 'an' were omitted entirely from the analysis on account of the large number of variant pronunciations to which they are subject" (p. 311). Given the very high frequency of these words, the reliability of French et al.'s results must be in doubt; this applies also to the re-analysis of their data by Tobias[35], even though he altered the accent of the transcriptions from that of New York City to what he claimed was GA even though he listed only one retroflex vowel, /ɝ/, and retained /ʍ/. A pioneering attempt to use spoken data from 30-month-old children[36,37] may be of historical interest to speech and language therapists, but not for present purposes.

Hayden[28] recorded and transcribed six lectures given in the *Orientation and English Language Program* for foreign students at the University of California; the phoneme tokens in the transcriptions totalled 65,122. Her Table 1 (pp. 220, 221) Hayden[28] gives data for 25 consonants and 14 vowels. Merging /ʍ/ with /w/ would produce the standard set of consonants, but the shortfall of vowels is problematic. None of the six retroflex vowels shown in **Table 3** above is listed, which is odd for data gathered in California, and there is no way of re-calculating her data to allow for this. Her data are therefore not used here.

Fowler[21] transcribed words yielding the first 5000 phonemes from each of three pieces of literary prose (also the whole of *Story of a Fierce Bad Rabbit* by Beatrix Potter, hardly relevant here). Because he too used only a 39-phoneme set, his tabulations are not considered here.

Delattre[38] analysed 2000 syllables taken equally from prose and drama, yielding between 6000 and 7000 phoneme tokens. However, according to Isengel'dina[39], his phoneme inventory consisted of 24 consonants and only 15 vowels — or 16, if /ʌ/, which is mentioned but not included in the frequency tables, is counted.

Carterette and Jones[40] produced a massive (645 pages) study of "informal spoken language" (p. 3) gathered from 24 students from junior college classes in a city in southern California. The transcribed conversations amounted to 15,694 words containing 48,708 phonemes (p. 14), but since they too used a 39-phoneme system, no use is made of these data here. They also refer (p. 15) to three prose passages but provide no information about phonetic transcriptions or outcomes from them.

Mines et al.[15] investigated phonemes in conversational English by using transcriptions of about 10 minutes of each of 26 tape-recorded interviews with adult subjects in California, 887 phonemes (p. 224). Yet their data cannot be used here. Their phonemic key (Table 1, p. 224) contains only one of the retroflex vowels listed in the GA section of **Table 3** above, and none of the related GB centring diphthongs, so there is no way of re-calculating their data to allow for this.

A listing of 'English sounds ranked by frequency. Most common sounds in spoken English have appeared on the internet[41], which gives this description: "This classification is based on data compiled using the Carnegie Mellon Pronouncing Dictionary correlated to a frequency list of the British National Corpus, i.e., American pronunciations with

British word-usage." The listing then shows 39 phonemes; that is, despite the phonetics being based on an American dictionary, it omits all six retroflex vowels. And yet again, there is no way to re-calculate.

It is telling that Wang and Crawford [22] and Gerber and Vertin [23] used only consonants in their comparative analyses of previous authors' data; they must have realised that those authors' data on vowels were deficient. Isengel'dina [39] provided a severe critique of many of the analyses mentioned so far, on the grounds of "the lack of a scientific phonemic approach to the material of their frequency counts" (p. 26); given what we too have noted about insufficient numbers of vowel phonemes analysed in several studies, the case seems unarguable.

Curiously, this winnowing leaves only the two earliest authorities discovered as providing reliable enough text frequency data to be tabulated, even though neither, strictly speaking, analysed the GA accent; also, despite their having been published about 150 and 100 years ago, respectively. The first person anywhere in the world to calculate phoneme frequencies for English appears to have been the philologist William Dwight Whitney [16,17]. He selected 10 literary passages, five poetic, five prose, from authors ranging from Shakespeare (Anthony's speech over Caesar's body), and Psalm 27 (King James version), through Milton, Gray's *Elegy* and Dr Johnson, to Tennyson and Macaulay. He transcribed the first 1000 phonemes of each selection in (mainly) his own pronunciation, using his own phonetic symbols, and took the figures derived from the 10,000 phonemes (from a total of 2726 words) to be a first approximation to a general statement of phoneme text frequencies.

Whitney's consonants are the usual 24, as listed in **Table 2**, except that he interprets words beginning <wh> as pronounced with /hw/ (and counts the two phonemes separately towards /h, w/). For his frequencies for consonants, see the left side of **Table 4**. Most of his 20 vowels (if the interpretations of his descriptions and symbols are correct — see the right side of **Table 4**) are also familiar, with the following exceptions. He counts syllabic /l̩ n̩/ (as in *battle, reckon*) as vowels, but not syllabic /m̩/ (as in *chasm, prism*), which would seem an obvious addition if syllabic consonants are to be counted separately.

Table 4. Phonemes and text frequencies in Whitney [16,17].

Consonants		**Vowels**		
IPA	%	Symbols Whitney's	IPA	%
r	7.44	Ĭ	ɪ	5.90
n	6.76	ə	ʌ, ə	5.66
t	5.93	E	ɛ	3.34
d	4.94	Œ	Æ	3.32
s	4.69	Ī	iː	2.80
l	3.84	Ă	ɒ	2.59
ð	3.83	Ū	uː	2.00
m	3.06	ɑi	aɪ	1.91
z	2.92	Ǝ	ɜː	1.85
v	2.37	Ō	oʊ	1.76
h	2.34	Ē	eɪ	1.61
w	2.31	A	ɔː	1.54
k	2.17	Au	aʊ	0.83
f	2.06	ɑ	ɑː	0.56
p	1.71	Æ	eə	0.47
b	1.64	Ŭ	ʊ	0.44
ʃ	0.86	L̩	əl	0.35
g	0.79	N̩	ən	0.16
ŋ	0.79	Ai	oɪ	0.12
j	0.66	Ŏ	O	0.08
Θ	0.58			
ʧ	0.53			
ʤ	0.47			
ʒ	0.02			
Totals	**62.71**			**37.29**

Note: All frequencies taken from table in Whitney [16] (p. 274).

Whitney describes his original Massachusetts accent as non-rhotic[16] but, in an apparent nod to more general American usage[16], interprets pre-consonantal letter <r> as representing a full /r/ phoneme, rather than retroflexion of the preceding vowel. As a consequence, he provides no separate account of retroflex vowels, though he alludes[16] to the schwa ending of the vocalic sounds in *care, fear, sore, cure, fire, sour*. He maintains he has 'short o' (/ɒ/) in, for example, *not, what, knowledge*. Being a New Englander, Whitney had both /ɑː/ and /ɔː/ in his accent, judging by his examples: *far, father, are, margin* vs *war, ball, law, dwarf*. This is unlike the current GA accent, where /ɔː/ is absent — see **Table 3** — having merged variously with /ɑː, oə/. In all these respects, his accent seems more like current GB than current GA.

He seems to merge /ʌ, ə/: his principal example is (unstressed?) *but*, but he gives copious examples which clearly (today) have /ə/, e.g. <a> in *woman, pagan*, <o> in *carol*, <e> in *absent*. Finally, he thinks his accent contains (rare) occurrences of a 'true short /o/'[16] (pp. 215, 216), by which he means, not the GB phoneme /ɒ/, but something akin to French /o/. He maintains he has such a vowel phoneme himself, in very few words, but enough to distinguish (for example) *none* /non/ from *known* /nəʊn/ - perhaps wishful thinking. For present purposes, his data for /o/ have been disregarded.

Based on that decision and on the identifications of Whitney's vowel symbol shown in the right-hand side of **Table 4**, that table shows the frequencies for his phonemes expressed as percentages. (Whitney gives no numbers for the frequencies of phonemes, only the percentages as reproduced above. However, the absolute numbers would be the numbers of the percentages in **Table 4** above, multiplied by 100.).

Nearly 50 years after Whitney, Dewey[18], which incidentally is speckled with his characteristic reformed spellings, e.g., *ar, descriptiv, difthong, sillable*, provided only the second set of frequencies of any sort ever. His corpus of 100,000 words containing 372,729 phonemes comprised selections from a wide variety of sources (newspaper editorials and articles, modern fiction and drama, speeches by Abraham Lincoln, Theodore Roosevelt and Woodrow Wilson, letters, advertizements, religious English (including the only pre-modern extract, from Mark's Gospel), etc.). All were transcribed over four years by four assistants using the Revised Scientific Alphabet, also known as the N.E.A. (National Educational Association) Phonetic Alphabet.

Dewey's Table 15[18] (p. 125), gives the frequencies of phonemes in the transcribed texts on a 48-phoneme basis based on the NEA alphabet. By merging the frequencies of two symbols each for /æ, ɑː, ɪ, juː/, and adding the combined values for the 2-phoneme sequence /juː/ to both plain /uː/ and /j/, it was possible to arrive at frequencies for a 43-phoneme set, and these are shown in **Table 5**. Adding the combined values for /juː/ into two places entailed raising the overall total of the percentages to 100.31%; though this is a slight breach of logic, the alternative (a complete recalculation) would have made so little difference to the figures that the effort needed would have been disproportionate. It should also be noted that Dewey's phonemes contain no retroflex vowels, and only one of the three GB-style centring diphthongs, /eə/ - the other two, /ɪə, ʊə/, are rare, so their absence would probably not greatly affect the frequencies shown. Since Dewey was based at Harvard, his data may therefore reflect a non-rhotic New England accent quite like Whitney's, since very few of his frequencies differ significantly from Whitney's — except that his value for /ɪ/, 8.53%, is much higher than Whitney's 5.90%, and he does not include /o/. An oddity of both these authors' data is that the schwa vowel /ə/ is in fifth place in overall rank orders – later analyses consistently show it as the most frequent phoneme in GB by some margin, and higher than fifth place in GA. Perhaps over-careful pronunciations were transcribed.

Table 5. Phonemes and text frequencies in Dewey[18].

Consonants		Vowels	
IPA	**%**	**IPA**	**%**
N	7.24	ɪ	8.53
T	7.13	ə	4.63
R	6.88	ɛ	3.44
S	4.55	Æ	3.72
D	4.31	ɒ	2.81
L	3.74	iː	2.12

Table 5. *Cont.*

Consonants		Vowels	
IPA	**%**	**IPA**	**%**
Đ	3.43	uː	1.91
Z	2.97	eɪ	1.84
M	2.78	ʌ	1.70
K	2.71	oʊ	1.63
V	2.28	aɪ	1.59
W	2.08	ɔː	1.26
P	2.04	ʊ	0.69
F	1.84	ɜː	0.63
H	1.81	aʊ	0.59
B	1.81	ɑː	0.49
Ɒ	0.96	eə	0.23
J	0.91	oɪ	0.09
ʃ	0.82		
G	0.74		
ʧ	0.52		
ʤ	0.44		
Θ	0.37		
ʒ	0.05		
Totals	**62.41**		**37.90**

Note: All frequencies derived from Dewey's Table 15 (p. 125)[18].

Apparently, the only person to have used Dewey's frequencies for further research was Zipf[42]. However, he used only Dewey's consonant data, and did not seek to provide new phoneme frequency estimates of his own. No further account is taken of this reference.

It is beginning to look as though the data we will provide on the GA accent will be the first reliable set of frequencies for GA, in the sense of being based on a justified 44-phoneme analysis which includes all six retroflex vowels and omits /ɒ, ɔː/.

1.6. Phoneme Text Frequencies in the General British Accent

For GB/RP, the only published sources for such frequencies are Fry[9], Denes[10,11] and Knowles[1]. (The eighth edition of *Gimson's Pronunciation of English*[6] provides tables giving figures described in footnotes as "conflated"; on inspection, they turn out to be the arithmetic means of Fry's and Knowles's, and are not used here.)

Fry[9] and the unnamed colleagues who helped him used: "… the conversational matter contained in Daniel Jones' *Phonetic Readings in English*. This consists of a succession of anecdotes expressed in fairly colloquial language. The transcription used in this book represents a typical Southern English pronunciation" (p. 104). Jones's book[43] was first published in 1912; the texts are given in both IPA and 'ordinary spelling.' It seems clear that the anecdotes were not first spoken, then written down, then transcribed into IPA; rather, the 'ordinary spelling' versions were the originals. Gramophone records of the texts were available, but Jones said of those[43]: "They are spoken by myself" (p. iv). Fry[9] reported that "The total number of sounds [phonemes] counted [in Jones's transcriptions] was just over 17,000" (p. 105).

Although based at the Bell Telephone Labs in New Jersey, Denes[10,11] also analysed GB/RP, using two sets of 'phonetic readers': Daniel Jones's (though in the 36th edition, 1959), thus creating considerable overlap with Fry's corpus, and a new set by Scott[44]. Scott's book contains 38 'phonetic texts' in IPA without 'ordinary spelling' versions; a few of the titles (transcribed from IPA into 'ordinary spelling') are '1. An appointment'; '15. Cousin James'; '35. Inland Revenue'. Scott gave no details of the sources of the texts — did he perhaps write them all? — but in the preface Daniel Jones wrote[43]: "[T]he texts are faithful representations of current Spoken English and free from unnatural or bookish expressions" (p. iv). Denes's corpus, drawn from both sources, amounted to 23,052 words containing 72,210 phonemes.

Knowles[1] described his corpus as follows: "Ten different types of text, five written and five spoken, ranging from a seed catalogue to a passage from *Pygmalion* to recorded interviews, were transcribed [into IPA representing RP]. The first 1000 phonemes of each text were counted, making a total of 10,000 phonemes" (p. 223). Bernard Shaw's play *Pygmalion* was first performed in 1913, so one of Knowles's sources was of the same vintage as all of Fry's and half of Denes's. It must be hoped that Knowles excluded Eliza Doolittle's utterances in broad Cockney, or that none featured in the first 1000 phonemes. No details of that or of Knowles's other texts are known — it would be particularly interesting to know more about the recorded interviews.

All three authors presented their results as lists of phonemes with percentages in decreasing order of frequency; all are presented (after correction) below. Corrections were needed because both Fry's and Knowles's lists, as originally published, contained errors. Fry's percentages added up to 98.81% rather than 100%, too large a discrepancy to arise from rounding errors. In a footnote in the second edition of his *Pronunciation of English,* Gimson[8] said:

> In the original article, an error arose in the figures for /t/, /d/, and /r/, resulting in a total discrepancy of 1.19%. These figures have been corrected by Mr G Perren (British Council, London), and the total discrepancy has been reduced to 0.01%. The list quoted [for consonant phonemes on p. 219; the list of vowel phonemes is on p. 148] includes the revised percentages for /t/, /d/, and /r/ (p. 219).

The corrected data have been used here.

In Knowles[1] (pp. 223, 224), the phonemes (consonants first, then vowels) were listed in decreasing order of frequency — except that /f/, at 0.66%, fell between /p/ at 2.05% and /v/ at 1.94%. Consequently, the total for all phonemes (not given by Knowles) was 98.63%, again too large a discrepancy to result from rounding errors. In response to a query, Alan Cruttenden, editor and reviser of several later editions of *Gimson's Pronunciation of English* (personal communication to Greg Brooks, provided a table giving Knowles's figures with the percentage for /f/ increased to 1.66%. This adjustment still left /f/ out of numerical order, and the overall percentage at 99.63%. Increasing the percentage for /f/ to 2.03% instead puts it in the correct place in the list and brings the overall total up to 100%; the figure of 2.03% has been used here.

The text frequency data from all three analyses of GB are shown in **Tables 6** (consonants) and **7** (vowels). The percentages shown are calculated across all phonemes. Denes's data, originally calculated to four places of decimals, have been rounded to two places; Fry's and Knowles's data were already shown to two places. The phonemes in each table are listed in decreasing order of frequency in Fry's data. All three authors used the same set of 44 phonemes, as listed in **Tables 2** and **3**. The right-hand column in each of these tables shows our new data on text frequencies at the 4000-word level.

All three sets of previous figures are very similar, with no startling changes in frequency or rank. The fact that the schwa vowel /ə/ (which in English normally occurs only in unstressed syllables) is by some distance the most frequent phoneme in GB reflects the strong reducing influence of word stress on vowel qualities in unstressed syllables, a key feature of British English phonology. (Consider, for example, the very different pronunciations of the word *laboratory*, four syllables with second syllable stress in GB, versus five syllables with fourth syllable stress in GA.) However, even the most recent of these references is nearly 40 years old – so the analyses must be due for updating.

Table 6. Consonant phonemes and text frequencies: new data and three previous analyses of the GB accent.

Phoneme	as in	Fry[9]	Denes[10,11]	Knowles[1]	Our Data
/n/	Now	7.58	7.08	7.65	7.38
/t/	Tie	6.42	8.40	7.48	7.60
/d/	Dye	5.14	4.18	4.12	3.61
/s/	Sue	4.81	5.09	4.77	4.41
/l/	Low	3.66	3.69	3.91	3.91
/ð/	This	3.56	2.99	3.37	3.86
/r/	Run	3.51	2.77	3.62	2.87
/m/	Moon	3.22	3.29	2.29	2.88

Table 6. *Cont.*

Phoneme	as in	Fry [9]	Denes [10,11]	Knowles [1]	Our Data
/k/	Cup	3.09	2.90	2.89	3.29
/w/	Well	2.81	2.57	2.53	1.98
/z/	Zoo	2.46	2.49	3.05	1.10
/v/	View	2.00	1.85	1.94	2.44
/b/	Book	1.97	2.08	2.17	2.71
/f/	Few	1.79	1.73	2.03	1.89
/p/	Pie	1.78	1.77	2.05	2.15
/h/	House	1.46	1.67	1.00	1.72
/ŋ/	Ring	1.15	1.24	0.94	0.60
/g/	Good	1.05	1.16	0.93	0.88
/ʃ/	Shoe	0.96	0.70	0.82	0.90
/j/	Yell, union	0.88	1.53	1.26	1.22
/ʤ/	Jam	0.60	0.51	0.63	0.58
/ʧ/	Chew	0.41	0.37	0.53	0.60
/θ/	Thin	0.37	0.60	0.57	0.45
/ʒ/	Genre	0.10	0.05	0.04	0.05
Total		**60.78**	**60.73**	**60.59**	**59.08**

Table 7. Vowel phonemes and text frequencies: new data and three previous analyses of the GB accent (all data shown are percentages).

Phoneme	as in	Fry [9]	Denes [10,11]	Knowles [1]	Our Data
/ə/	*A*bout	10.74	9.04	10.49	8.50
/ɪ/	Ink	8.33	8.25	8.26	6.27
/e/	End	2.97	2.81	2.57	2.32
/aɪ/	Ice	1.83	2.85	2.22	1.89
/ʌ/	Up	1.75	1.67	1.41	1.60
/eɪ/	Aim	1.71	1.50	1.54	1.80
/iː/	Eve	1.65	1.79	1.80	4.23
/əʊ/	Owe	1.51	1.75	1.59	1.28
/æ/	Ash	1.45	1.53	1.80	3.38
/ɒ/	Ox	1.37	1.53	1.73	2.56
/ɔː/	Awe	1.24	1.20	1.36	1.45
/uː/	Ooze	1.13	1.42	1.46	2.53
/ʊ/	Pull	0.86	0.77	0.38	0.43
/ɑː/	Art	0.79	0.78	0.56	0.57
/aʊ/	Ouch	0.61	0.77	0.65	0.62
/ɜː/	Err	0.52	0.67	0.62	0.64
/eə/	Air	0.34	0.43	0.31	0.42
/ɪə/	Ear	0.21	0.29	0.36	0.25
/ɔɪ/	Oink	0.14	0.09	0.26	0.09
/ʊə/	Sure	0.06	0.14	0.04	0.06
Total		39.21	39.27	39.41	40.89
GRAND TOTAL		**99.99**	**100.00**	**100.00**	**100.00**

1.7. First Results: Some Changes in the General British Accent Over Time

To investigate changes in the GB accent over time, we compared the new data shown in **Tables 6** and **7** to the arithmetic means of the three earlier analyses shown there. We set a criterion of a difference of at least 1 percentage point between the two figures, and this yielded the various possible changes which we proceeded to analyse. Some differences appear to indicate actual changes in the GB accent over time; others cannot, to our knowledge, be aligned with trends recorded in published commentaries.

1.7.1. Vowels

One of the most interesting shifts in phonemes over time is the decline in /ɪ/ from an average frequency of 8.28%

for the previous frequency counts to 6.27% in our data. A shift of just over 2 percentage points. In contrast, the phoneme /i:/ has increased from an average frequency of 1.75% to 4.23% in this most recent count: a shift of almost 2.5 percentage points. We believe that these shifts are connected and consistent with published commentary and data reaching back almost 80 years. Ramsaran[45] provides an analysis of a small data set collected, during the 1970s drawing from the RP speech of three age groups. Her focus is on the shift from /ɪ/ to /i:/ preconsonantally (in words such as *happiness, dutiful,* etc.). She states that the figures offer some evidence that[45] "… in a few representative individuals the ratio of preconsonantal /i:/ to /ɪ/ is increasing (the older speaker exhibiting a ratio of 5:1 and the younger ones 9:1 and 19:1)" (p. 185). For example, Wells[46,47] and Windsor Lewis[48] also acknowledge this strong tendency for /i:/ to replace /ɪ/. Knowles[1] also touches upon this shift. Cruttenden[6] emphatically states that the presence of "a vowel nearer in quality to /i:/, rather than /ɪ/ is now the norm in GB, finally in words like 'lady, sloppy, happy, donkey, prairie…" (p. 113). Carney[3] appears conflicted about the phenomenon. In several places in *A Survey of British Spelling*, he states that /ɪ/ occurs word-finally, and when it does, it is predominantly spelt with <y> (e.g., 135, 139, 161, 380, 430). However, in the same publication[3] (pp. 134, 135), he also acknowledges that many of what he calls younger RP speakers no longer use /ɪ/ word finally. He notes that a short form of /i:/ features instead. The next largest difference between our data and the average of the earlier phoneme frequency counts itemised in **Table 7** is in /ə/, the schwa: a negative difference of 1.59 percentage points. However, the previous counts are discrepant: the difference between Fry's count and Denes's is 1.7 percentage points, and between Denes's and Knowles's is 1.45 percentage points. These discrepancies may render any comparison of our data with the previous frequency counts' average value tenuous. It is therefore difficult to be sure if the differences reported record real shifts across time or different choices at the stage of transcription. Cruttenden[6] notes that variation in patterns of accentuation for particular words can occur because of rhythmic and analogical pressures. Such pressures result in changes in the quality of vowels. Some suffixes, for example {-able}, can vacillate between accent neutral status (e.g., *question* and *questionable*) or feature transfer of accent to the first syllable (e.g., *admire* versus *admirable*). Some of these variations in accentual patterns can lead to concurrent alternative pronunciations: alternatives that need to be chosen between (e.g., *kilometre* as /ˈkɪləmi:tə/ or /kɪ ˈlɒmɪtə/) at the transcription stage. Word accentual instability may therefore play a part in the percentage point differences across the four data sets. Cruttenden[6] and Knowles[1] note that grammatical (function) words such as prepositions, pronouns, articles, etc., have accented and unaccented (weak) forms. If some of these were given what Brooks[5] calls their 'full pronunciation' (p. 60) at the transcription stage, rather than their more usual weak form, as often found in connected text, then the count for /ə/ would also be reduced and the count for other vowels would have risen accordingly.

Reasons for the increase of 1.79 percentage points between our new data and the average of the previous counts for the phoneme /æ/ are harder to speculate upon. Word accentual factors could have played a small role. However, it is also the case that some words which used to have /ɑ:/ are now universally pronounced with /æ/, for example, *plastic* and the last syllable of *aftermath*.

Two other differences to note between our data and the average of the previous counts of vowel phonemes arise for/ɒ/ and /u:/. Just over a 1 percentage point increase in both features in the new data set. Again, we cannot tie these to any published commentary on trends and therefore offer no firm reasons for the rises. However, the rise in /ɒ/ might be due to the much higher frequency in the corpus of technological words such as *rocket, blog, podcast*, etc. and the word *technology* itself, in recent decades.

1.7.2. Consonants

The average frequency count for /z/ across the three previous frequency counts is 2.67%. In our data /z/, has a frequency count of 1.1% — a decrease of 1.57 percentage points. We cannot map this drop to any recent or documented changes in the GB accent. Part of the decrease could be explained by a lower number of singular verbs and plural nouns featuring {-es} or {-s} realised as /z/ in our selection from the corpus. This is speculative. There was broad agreement across all four data sets for the other consonants in GB.

The size of the sample upon which frequency data is based can skew findings. Knowles himself[1] concedes this point when he states that his sample size is "probably too small to prevent some distortion of this kind" (p. 224) — 'distortion'

meaning the inflated frequency of the key words in texts.

1.8. Lexical Frequencies

The earliest study of GB/RP (Trnka[25]) and one mid-century US study (Hanna et al.[7]) were based on dictionaries or word lists, and thus resulted in lexical frequencies. (Another US author, Roberts[49], used a corpus from a dictionary, and for an attempt to calculate lexical phoneme frequencies, used a subset of 66,534 phoneme tokens. However, his phoneme inventory contained only eight vowels, and his statistics for vowel phonemes therefore, cannot be reliable. This study, like so many mentioned above, is not used here).

Bohumil Trnka was based at Charles University in Prague throughout his career, and was a co-founder, the first secretary, and later the leader of the Linguistic Circle of Prague. His monograph *A Phonological Analysis of Present-day Standard English*[25] (1935; revised edition published in Japan in 1966 and in the US in 1966[50] and 1968, as chronicled by Dušková[51] contains highly detailed analyses of phoneme occurrences in monomorphemic words containing no more than two vowel phonemes (and therefore no more than two syllables). In the preface[25,50] (p. 2 and p. v respectively), he describes his corpus as follows: "Nearly all the word material tabulated in this work is taken from the *Pocket Oxford Dictionary of Current English*" (F. Fowler and H. Fowler[52]). The total number of words he analysed appears to be 5654, comprising 3203 monosyllables (figure inferred from Trnka[50], p. 113), 2221 disyllables stressed on the first syllable[50], pages 122 and 230 disyllables stressed on the second syllable[50] (p. 133). For reasons that will become apparent, figures for the number of phonemes he analysed have a margin of uncertainty.

In the opening of the Preface to the later edition (p. i), Trnka[50] says: "The part of the book devoted to the statistical investigation of productivity [equals frequency] of phonemes in the formation of monomorphemic monosyllables and disyllables did not undergo many changes. It was revised carefully, but no additions and modifications were introduced in the word lists and statistical tables." For present purposes, therefore, the rest of this analysis is based solely on the 1966 edition.

Trnka[50] gives overall numbers for 41 of the 44 phonemes of RP/GB, the exceptions being /ʊə, ɔɪ, ə/, plus the 'triphthongs' /aɪə, aʊə/ (as Trnka analyses, for example, *fire, flower*). Trnka[50] discusses /ʊə/ but does include it in his inventory of phonemes — but it is rare (so its absence barely affects the overall picture). It is clear it was already disappearing by 1966 since on page 17 of that edition Trnka transcribes the word *sure* as both /ʃ ʊə/ and /ʃ ɔɪ/ (the 1935 edition shows no trace of this).

On /ɔɪ/ he says[50] "[T]his diphthong represents the combination of two contiguous phonemes /ɔ/+/j/," (p. 17) and this is how he represents and analyses this sequence throughout. Here, however, since we treat the second element as /ɪ/, a figure of 40 for /ɔɪ/ has been reached by counting all the words containing it that Trnka mentions (a few more disyllables with second-syllable stress may have been missed through having been listed in a section of the original 1935 edition which was omitted from the revised edition, see the preface[5]); this number has also been deducted from his figures for /ɔ, j/.

We consider 'triphthongs' to be sequences of two syllables, the second being /ə/. Accordingly, we have dissolved /aɪə, aʊə/ into /aɪ+ə, aʊ+ə/, and added the figures for /aɪə, aʊə/ to those for /aɪ, aʊ/. Arriving at a figure for /ə/ was more problematic. The occurrences 'snipped off' /aɪə, aʊə/ total 73, far too few to represent what is widely known to be the most frequent phoneme in RP/GB. We instead reasoned that the figure for /ə/ must be greater than that for Trnka's otherwise most frequent phoneme (/t/, N = 1659), and that the great majority of Trnka's disyllables would have had /ə/ in the unstressed syllable, so took his overall figure for disyllables of 2451, and arrived at the 'educated guess' of 2000 for /ə/.

On those assumptions and adjustments, the figures shown in **Table 8** were arrived at, and from those data frequencies (percentages) and ranks were calculated for 43 phonemes. One further caveat: since Trnka used only monomorphemic mono- and disyllables, an analysis incorporating polymorphemic mono- and disyllables and words of more than two syllables might well lead to rather different estimates.

Hanna et al.[7] was an early, and massive, attempt to apply computer technology to the analysis of (American) English pronunciation and spelling, and was highly influential in the development of phonics teaching materials — see Cochrane and Brooks[53]. They used a corpus of 17,310 words, the majority of which (15,284) were derived from the *Teacher's*

Word Book of 30,000 Words, Part I (Thorndike and Lorge [54]), which contained 19,440 entries after eliminating slang, foreign words, proper names, abbreviations, etc. (the categories of exclusions are listed on p. 12 of the report); Hanna et al. then added 2026 words from *Webster's New Collegiate Dictionary*, 6th edition (1961). The transcriptions of the words' pronunciations were all taken from that dictionary, based on its pronunciation key; Hanna et al. [7] reasoned that this "provided the kind of general American-English 'dialect' most suitable for the proposed phonological analysis of the orthography." (p. 13). This rather vague statement does not fully specify the accent codified in the *New Collegiate Dictionary*'s pronunciation key, which differs in various respects from current GA, as shown below. Because the computers of the day did not provide IPA symbols, Hanna et al. used ASCII-character-set codes for phonemes; for example, /ʊ/ was O7.

Table 8. Lexical GB/RP phoneme frequencies in monomorphemic monosyllables and disyllables based on Trnka [50].

Consonants				**Vowels**			
IPA	**N**	**%**	**rank**	**IPA**	**N**	**%**	**rank**
n	1354	5.96	5	ɪ	677	2.98	13
t	1659	7.31	2	ə	2000	8.81	1
r	1045	4.60	7	E	521	2.29	16
s	1563	6.88	4	Æ	746	3.29	10
d	830	3.66	9	ɒ	441	1.94	17
l	1633	7.19	3	iː	319	1.41	21
ð	64	0.28	38	uː	276	1.22	27=
z	203	0.89	33	eɪ	365	1.61	19
m	715	3.15	11	ʌ	526	2.32	15
k	1276	5.62	6	əʊ	317	1.40	22
v	305	1.34	24	aɪ	359	1.58	20
w	315	1.39	23	ɔː	284	1.25	25
p	909	4.00	8	ʊ	42	0.18	40
f	550	2.42	14	ɜː	221	0.97	32
h	240	1.06	30	aʊ	135	0.59	36
b	704	3.10	12	ɑː	276	1.22	27=
ŋ	188	0.83	34	eə	33	0.15	42
j	172	0.76	35	ɔɪ	40	0.18	41
ʃ	283	1.25	26	ɪə	49	0.22	39
g	405	1.78	18				
ʧ	239	1.05	31				
ʤ	275	1.21	29				
θ	119	0.52	37				
ʒ	31	0.14	43				
Subtotals	**15,077**	**66.41**			**7627**	**33.59**	
Grand Total				**22,704**			

Note: "=" indicates that there are two or more phonemes with an equal ranking.

Hanna et al.'s Tables 9 and 10 [7] provide estimated lexical frequencies and percentages of the vowel and consonant phonemes, respectively, in the implied US accent. The (quasi-) phonemes include /ʍ/ as in a conservative pronunciation of words like *what*, syllabic /l̩, m̩, n̩/ as in *table, prism, Haydn*, and the 2-phoneme sequences /ks, kw, juː/ as in *fox, quick, union*. But there are, for no reason we can discern, two entries for /ɒ/, and the treatment of possible retroflex vowels is inconsistent (especially for researchers based in California): /ʊɚ/ is missing (the example word *sure* is listed under /ʊ/, see page 25); there are two entries for retroflex schwas (apparently for stressed and unstressed syllables separately); /ɑɚ, oɚ/ are subsumed into /ɑː, ɔː/ respectively; and only /ɛɚ, ɪɚ/ are handled as in current GA.

Because of those uncertainties, for this article, Hanna et al.'s 1966 data [7] have been re-analysed. The frequencies for the two entries each for /ɒ/ and schwa have been merged; that for /ʍ/ has been added to both /h/ and /w/ as this seems to be how Hanna et al. [7] thought of it; those for syllabic /l̩, m̩, n̩/ have been treated as sequences of schwa plus plain /l,

m, n/ and added to both those consonants and /ə/; and the figures for the 2-phoneme sequences /ks, kw, juː/ have been added to those for each of the constituent phonemes. This process increased the total number of phoneme occurrences by 2.3% — see the foot of **Table 9**. **Table 9** shows all the phonetic items in both Hanna et al.'s[7] ASCII-character-set codes (for cross-checking with their report) and IPA, their original and our revised frequencies, and a new rank order covering all the resulting 43 phonemes (there was no basis for calculating a frequency for /ʊɚ/, which would in any case have been very small and would not have affected the overall picture).

Table 9. Original and re-calculated lexical frequencies of phonemes for the US accent analysed by Hanna et al.[7].

Consonants						**Vowels**					
Symbols		**N**		**Revised**		**Symbols**		**N**		**Revised**	
Hanna	**IPA**	**Hanna**	**Revised**	**%**	**rank**	**Hanna**	**IPA**	**Hanna**	**Revised**	**%**	**rank**
N	n	7662	7790	7.01	4	I3	ɪ	7815	7815	7.03	2
T	t	7796	7796	7.02	3	ə	ə	6013	6900	6.21	5
R	r	9304	9304	8.37	1	E3	ɛ	3646	3646	3.28	11
S	s	6328	6599	5.94	6	A3	æ	4340	4340	3.91	9
D	d	3703	3703	3.33	10	O3, O5	ɒ	1789	1789	1.61	20
L	l	5389	6051	5.45	7	E	iː	2538	2538	2.28	16
T2	ð	149	149	0.13	41=	O6	uː	453	1641	1.48	21
Z	z	997	997	0.90	28	A	eɪ	2248	2248	2.02	18
M	m	3503	3600	3.24	12	U3	ʌ	1410	1410	1.27	25
K	k	4714	5181	4.66	8	O	oʊ	2587	2587	2.33	15
V	v	1492	1492	1.34	23	I	aɪ	1482	1482	1.33	24
W	w	626	902	0.81	30	O2	ɔː	767	767	0.69	32
P	p	3455	3455	3.11	13	O7	ʊ	368	368	0.33	38
F	f	2022	2022	1.82	19	E5, U2	ɝ	2957	2957	2.66	14
H	h	778	858	0.77	31	OU	aʊ	406	406	0.37	37
B	b	2306	2306	2.08	17	A5	ɑː	580	580	0.52	34
NG	ŋ	616	616	0.55	33	A2	eɚ	220	220	0.20	39
Y	j	120	1308	1.18	27	OI	oɪ	149	149	0.13	41=
SH	ʃ	1537	1537	1.38	22	E2	ɪɚ	198	198	0.18	40
G	g	1342	1342	1.21	26	U	juː	1188			
CH	ʧ	564	564	0.51	35						
J	ʤ	982	982	0.88	29						
T1	θ	411	411	0.37	36						
ZH	ʒ	102	102	0.09	43						
HW	ʍ	80									
KS	ks	271									
KW	kw	196									
L1	l̩	662									
M1	m̩	97									
N1	n̩	128									
Totals		**67,431**	**69,067**					**41,154**	**42,041**		
Grand totals		**Hanna**		**108,585**		**Revised**		**111,108**			
				Increase		**2523**		**(2.3%)**			

Note: "=" indicates that there are two or more phonemes with an equal ranking.

As already pointed out in Cochrane and Brooks[53], Hanna et al.'s[7] procedure produced some odd results. First, /r/ being in the first rank was based on counting every occurrence of letter <r> in their database as an instance of phoneme /r/ — but the great majority of its occurrences are in post-vocalic position, and should, therefore, have been analysed as parts of graphemes representing phonemes other than /r/. Secondly, <l> also features in some digraphs where it should not have been counted separately towards the frequency of /l/, e.g., in words like *walk*. Thirdly, their analysis implies that the accent they analysed contained the short vowel phoneme ɒ, despite this appearing to be lacking in all contemporary

and recent US accents.

2. Methods Used to Obtain New Estimates

Having whittled the long list of previous sources down to the seven we consider reliable and presented our analyses of those authors' data, we now proceed to present ours.

2.1. Method

2.1.1. Materials

The new analyses reported here are based on the Corpus of Contemporary American English (COCA)[55]. An acknowledged source that has been purchased online and granted permission for non-profit use with a limited sharing scope, which is provided by Mark Davies. Each distributed list includes a unique footprint to ensure proper usage tracking. The corpus contains more than one billion words of text (25+ million words each year, 1990–2019) from eight genres: spoken, fiction, popular magazines, newspapers, academic journals, and (with the March 2020 update): TV and movie subtitles, blogs, and other web pages in the GloWbE corpus. We have used a subset containing the 20,200 most frequent words in the full corpus, which considers separate entries for lemmas that have different parts of speech[56]. Inspection of the first 4000 words found them to be equally applicable to British English.

2.1.2. Transcriptions

The first author created parallel files for British English and American English; for the former, some spellings were changed in accordance with British usage. The second and third authors created the phoneme keys shown in **Tables 2** and **3**, which the first author then used, along with the decisions and features mentioned above, **Table 3**, to create IPA transcriptions of the 20,200 words in both GB and GA. Over several iterations, the transcriptions of the first 4000 words were checked by the second and third authors, and any general findings were applied to the full database, until authors 1−3 were satisfied that they represented the words' pronunciations to a satisfactory level of accuracy.

2.1.3. Analysis

The first author then used the Python programming language to calculate phoneme frequencies — his detailed program is presented in the annex/associated material and can be obtained on request to him at the given email address in the paper. We had divided the analysis into four levels iteratively on the first 1000, 2000, 3000 and 4000 words, and not on the full 20,200 words or on any subset beyond the first 4000 words. This is because we found that the phoneme frequencies stabilise at the 3000- or 4000-word level. We accounted for the fact that different parts of speech are treated as separate entries for the same lemma. Additionally, the COCA list provides the frequency of each lemma within the 20,200-word corpus.

The first author then used his program to count (1) the lexical frequencies of the phonemes in each accent separately, that is, by counting each word and its constituent phonemes separately only once; (2) the text frequencies of the phonemes in each accent separately. The latter involved multiplying the phoneme occurrences in each word by the frequency of the word in the full database. For example, the phonemes /ð, ə/ in *the* were each counted 22,038,906 times because that is the number of occurrences of *the* in COCA.

2.1.4. Procedures and Calculations

This procedure had been applied in the same way for each accent, taking into account the differences in the phoneme entries with their respective IPA transcriptions.

2.1.5. Lexical Frequencies

First, we need [Y] to count each phoneme only once per word at each level of analysis.
Second, we need [P] to calculate the total number of occurrences of the 44 phonemes [P_1 to P_{44}] or [P_1 to P_{45}] if we include /juː/, for each accent at each level of analysis.

$$\sum_{1}^{n} P_1 + P_2 + P_n = P(\#)$$

A percentage is then calculated:

$$\frac{[Y]\#\ Single\ occurrences\ of\ a\ specific\ phoneme\ per\ word\ sample\ in\ the\ dataset}{[P]\ Total\ count\ of\ all\ occurrences\ of\ the\ 44\ phonemes\ within\ the\ dataset\ of\ words} \times 100\%$$
$$= [Y]\%\ The\ relative\ frequency\ (percentage)\ of\ a\ specific\ phoneme\ occurring\ in\ the\ word\ data$$

Third, we need to sum all the resulting frequencies (%) of each phoneme to ensure they add up to 100%, thereby confirming the accuracy of the calculations.

2.1.6. Text Frequencies

First, we need [Z] to count all occurrences of each phoneme per word at each level of analysis.
Second, we need to multiply each [Z] value by the corresponding word frequency provided by COCA for each constituent word at each level of analysis.
Third, we need [P″] to sum the total values of all occurrences of the 44 (or 45) phonemes, as weighted by their respective word frequencies.

$$\sum_{1}^{n} P^{n}{}_{1} + P^{n}{}_{2} + P^{n}{}_{n} = P''(\#)$$

A percentage is then calculated:

$$\frac{\text{[Z]\#All occurrences of a specific phoneme per word sample in the dataset} \times \text{Frequency of each related word}}{\text{[P”] The total sum of all 44 phoneme occurrences} \times \text{Frequency of each related word within the dataset of words}} \times 100\%$$
$$= \text{[Z]\%The relative frequency (percentage) of a specific phoneme occurring in the word data}$$

Fourth, we need to sum all the resulting frequencies (%) of each phoneme to ensure they add up to 100% or thereabouts, confirming the accuracy of the calculations.

3. Results

The results for the first 1000, 2000, 3000 and 4000 words are shown below. NB: From this point onwards, column totals which are not exactly 100.00% contain variance arising from the rounding process.

The lexical and text frequencies for GB are in **Tables 10** and **11** respectively.

The lexical and text frequencies for GA are in **Tables 12** and **13** respectively.

These tables show that, in all cases, the frequencies do indeed change somewhat after the first 1000 and 2000 words but then stabilise, thus justifying our decision not to proceed beyond the 4000-word level. This is consistent with other indications in the literature, especially the small sample size validated in an analysis of Australian Aboriginal phonemes[57]. Coralie Cram and Claire Bowern also concluded that their results tentatively indicate a high level of validity for small datasets[58].

That there are also, as predicted by theory and expected, some significant differences between the two kinds of frequency, at least for certain phonemes, is shown more clearly in **Table 14**, in which the figures for both kinds of frequency in both accents at the 4000-word level are presented. In the table, the IPA symbols for GB phonemes are on the left, those for the nearest GA equivalents on the right. Note in particular that GB /ɒ/ is paired with GA /ɑː/, and that GB /ɑː/ is much further down the table than GA /ɑː/.

Table 10. Lexical frequencies of GB phonemes at 4 levels.

Consonants GB	LF1000	LF2000	LF3000	LF4000
/n/	6.97	7.26	7.49	7.50
/t/	6.97	7.55	7.58	7.42
/s/	6.14	6.07	6.13	6.10
/l/	5.69	5.55	5.65	5.54
/k/	4.25	4.74	4.73	4.97
/r/	4.20	4.55	4.69	4.80
/p/	3.29	3.31	3.44	3.49
/d/	3.77	3.54	3.36	3.40
/m/	3.22	3.21	3.01	3.11
/f/	2.26	2.21	2.00	1.98
/b/	1.78	1.76	1.67	1.67
/v/	1.55	1.57	1.67	1.65
/ʃ/	1.05	1.29	1.51	1.59
/j/	1.26	1.38	1.38	1.30
/w/	1.62	1.30	1.27	1.21
/z/	1.14	1.03	1.06	1.08

Table 10. *Cont.*

Consonants GB	LF1000	LF2000	LF3000	LF4000
/g/	0.98	0.96	0.92	0.99
/dʒ/	0.94	1.03	0.99	0.95
/ŋ/	0.75	0.77	0.81	0.83
/tʃ/	0.91	0.86	0.79	0.74
/h/	1.05	0.80	0.68	0.69
/θ/	0.62	0.56	0.42	0.40
/ð/	0.78	0.43	0.37	0.29
/ʒ/	0.09	0.10	0.12	0.12
Subtotals	**61.28**	**61.83**	**61.74**	**61.82**
Vowels GB	**LF1000**	**LF2000**	**LF3000**	**LF4000**
/ə/	8.04	8.50	9.14	9.23
/ɪ/	6.53	7.13	7.42	7.75
/iː/	3.59	3.56	3.38	3.20
/e/	3.36	3.29	3.24	3.18
/eɪ/	2.08	2.00	2.04	2.12
/æ/	1.60	1.58	1.76	1.86
/aɪ/	1.94	1.77	1.82	1.74
/ɒ/	1.64	1.52	1.49	1.54
/ʌ/	1.87	1.51	1.31	1.31
/əʊ/	1.39	1.42	1.29	1.22
/uː/	1.42	1.33	1.26	1.15
/ɔː/	1.55	1.17	1.02	0.98
/ɑː/	0.94	0.88	0.84	0.78
/ɜː/	0.73	0.77	0.74	0.71
/aʊ/	0.57	0.51	0.43	0.43
/eə/	0.41	0.41	0.37	0.32
/ʊ/	0.34	0.28	0.24	0.24
/ɪə/	0.37	0.30	0.25	0.22
/ɔɪ/	0.25	0.14	0.13	0.11
/ʊə/	0.09	0.09	0.11	0.09
Subtotals	**38.71**	**38.16**	**38.28**	**38.18**
Totals	**99.99**	**99.99**	**100.02**	**100**

Table 11. Text frequencies of GB phonemes at 4 levels.

Consonants GB	TF1000	TF2000	TF3000	TF4000
/t/	7.56	7.64	7.65	7.60
/n/	7.28	7.33	7.39	7.38
/s/	3.82	4.16	4.34	4.41
/l/	3.40	3.68	3.86	3.91
/ð/	5.13	4.38	4.06	3.86
/d/	3.73	3.67	3.63	3.61
/k/	2.67	3.04	3.18	3.29
/m/	2.82	2.86	2.86	2.88
/r/	2.19	2.57	2.77	2.87
/b/	3.07	2.87	2.49	2.71
/v/	2.70	2.54	2.49	2.44
/p/	1.71	1.94	2.09	2.15
/w/	2.30	2.10	2.03	1.98
/f/	1.87	1.91	1.90	1.89
/h/	2.11	1.88	1.78	1.72
/j/	1.15	1.21	1.23	1.22
/z/	1.13	1.10	1.11	1.10
/ʃ/	0.63	0.76	0.85	0.90

Table 11. *Cont.*

Consonants GB	TF1000	TF2000	TF3000	TF4000
/g/	0.84	0.86	0.86	0.88
/ʧ/	0.56	0.60	0.61	0.60
/ŋ/	0.53	0.56	0.59	0.60
/ʤ/	0.44	0.54	0.57	0.58
/θ/	0.47	0.48	0.45	0.45
/ʒ/	0.03	0.04	0.05	0.05
Subtotals	**58.14**	**58.72**	**58.84**	**59.08**
Vowels GB	**TF1000**	**TF2000**	**TF3000**	**TF4000**
/ə/	8.19	8.30	8.48	8.50
/ɪ/	5.60	6.00	6.17	6.27
/iː/	4.56	4.41	4.32	4.23
/æ/	3.93	3.57	3.46	3.38
/ɒ/	2.94	2.71	2.62	2.56
/uː/	3.01	2.74	2.62	2.53
/e/	2.04	2.21	2.29	2.32
/aɪ/	1.97	1.91	1.91	1.89
/eɪ/	1.71	1.74	1.78	1.80
/ʌ/	1.75	1.67	1.62	1.60
/ɔː/	1.66	1.54	1.48	1.45
/əʊ/	1.29	1.31	1.30	1.28
/ɜː/	0.61	0.64	0.64	0.64
/aʊ/	0.70	0.66	0.64	0.62
/ɑː/	0.50	0.55	0.57	0.57
/ʊ/	0.50	0.46	0.44	0.43
/eə/	0.45	0.44	0.43	0.42
/ɪə/	0.27	0.26	0.26	0.25
/ɔɪ/	0.10	0.10	0.10	0.09
/ʊə/	0.05	0.06	0.07	0.06
Subtotals	**41.83**	**41.28**	**41.2**	**40.89**
Totals	**99.97**	**100**	**100.04**	**99.97**

Table 12. Lexical frequencies of GA phonemes at 4 levels.

Consonants GA	LF1000	LF2000	LF3000	LF4000
/n/	6.99	7.26	7.49	7.47
/t/	7.05	7.59	7.64	7.81
/s/	6.18	6.08	6.15	6.09
/l/	5.70	5.56	5.65	5.53
/k/	4.26	4.73	4.74	4.96
/r/	4.21	4.55	4.71	4.80
/p/	3.30	3.31	3.45	3.48
/d/	3.85	3.60	3.40	3.42
/m/	3.23	3.21	3.01	3.11
/f/	2.27	2.21	2.01	1.98
/b/	1.79	1.78	1.68	1.68
/v/	1.56	1.57	1.67	1.65
/ʃ/	1.05	1.30	1.50	1.58
/j/	1.26	1.30	1.30	1.23
/w/	1.63	1.30	1.27	1.20
/z/	1.12	1.02	1.07	1.07
/g/	1.01	0.97	0.93	0.99
/ʤ/	0.87	0.97	0.95	0.91
/ŋ/	0.76	0.77	0.81	0.84
/ʧ/	0.85	0.80	0.73	0.70

Table 12. *Cont.*

Consonants GA	**LF1000**	**LF2000**	**LF3000**	**LF4000**
/h/	1.05	0.80	0.68	0.68
/θ/	0.62	0.56	0.42	0.40
/ð/	0.78	0.43	0.37	0.29
/ʒ/	0.09	0.10	0.12	0.12
Subtotals	**61.48**	**61.77**	**61.75**	**61.99**
Vowels GA	**LF1000**	**LF2000**	**LF3000**	**LF4000**
/ə/	5.98	6.67	7.40	7.52
/ɪ/	6.21	7.01	7.18	7.45
/iː/	3.57	3.54	3.38	3.19
/ɛ/	3.39	3.29	3.23	3.16
/ɚ/	3.21	2.96	2.85	2.76
/æ/	1.95	1.96	2.14	2.17
/ɛɚ/	0.48	0.47	0.44	0.39
/aɪ/	1.86	1.73	1.79	1.71
/ɑː/	2.11	1.90	1.84	1.84
/oʊ/	1.37	1.42	1.29	1.21
/ʊ/	0.32	0.23	0.19	0.18
/eɪ/	2.06	1.98	2.01	2.09
/oɪ/	0.23	0.13	0.12	0.11
/uː/	1.42	1.32	1.24	1.13
/aʊ/	0.57	0.51	0.43	0.43
/ʊɚ/	0.09	0.09	0.11	0.09
/ʌ/	1.88	1.50	1.30	1.30
/oɚ/	0.98	0.77	0.68	0.70
/ɑɚ/	0.48	0.44	0.40	0.36
/ɪɚ/	0.37	0.29	0.25	0.22
Subtotals	**38.53**	**38.21**	**38.27**	**38.01**
Totals	**100.01**	**99.98**	**100.02**	**100**

Table 13. Text frequencies of GA phonemes at 4 levels.

Consonants GA	**TF1000**	**TF2000**	**TF3000**	**TF4000**
/p/	1.71	1.94	2.08	2.16
/b/	3.07	2.88	2.77	2.72
/m/	2.82	2.86	2.85	2.88
/f/	1.87	1.91	1.89	1.90
/v/	2.70	2.54	2.48	2.44
/θ/	0.47	0.48	0.45	0.45
/ð/	5.13	4.38	4.05	3.86
/t/	7.61	7.68	7.68	7.65
/d/	3.77	3.71	3.66	3.65
/s/	3.83	4.17	4.34	4.42
/z/	1.12	1.10	1.10	1.10
/n/	7.28	7.33	7.38	7.39
/r/	2.19	2.57	2.76	2.88
/l/	3.40	3.68	3.85	3.92
/ʃ/	0.63	0.76	0.85	0.90
/ʒ/	0.03	0.04	0.05	0.05
/ʧ/	0.52	0.55	0.56	0.56
/ʤ/	0.40	0.50	0.53	0.54
/j/	1.15	1.19	1.20	1.19
/w/	2.30	2.10	2.03	1.98
/k/	2.67	3.04	3.18	3.30
/g/	0.85	0.87	0.87	0.88

Table 13. *Cont.*

Consonants GA	TF1000	TF2000	TF3000	TF4000
/ŋ/	0.53	0.56	0.59	0.60
/h/	2.11	1.88	1.77	1.72
Subtotals	**58.16**	**58.71**	**58.98**	**59.14**
Vowels GA	**TF1000**	**TF2000**	**TF3000**	**TF4000**
/iː/	4.56	4.40	4.30	4.23
/ɪ/	5.52	5.91	5.97	6.08
/ɛ/	2.05	2.22	2.29	2.32
/æ/	4.11	3.78	3.68	3.62
/ə/	7.11	7.14	7.27	7.30
/uː/	3.01	2.74	2.61	2.53
/ʊ/	0.50	0.44	0.42	0.40
/ʌ/	1.75	1.67	1.62	1.60
/ɑː/	3.39	3.14	3.03	2.98
/eɪ/	1.71	1.74	1.77	1.79
/oɪ/	0.10	0.09	0.09	0.09
/aɪ/	1.93	1.88	1.88	1.86
/ɚ/	1.90	2.03	2.08	2.10
/ɑɚ/	0.25	0.27	0.28	0.26
/ɛɚ/	0.48	0.47	0.46	0.45
/ɪɚ/	0.27	0.26	0.25	0.25
/oɚ/	1.17	1.08	1.04	1.03
/ʊɚ/	0.05	0.06	0.07	0.06
/oʊ/	1.28	1.31	1.29	1.28
/aʊ/	0.70	0.66	0.63	0.63
Subtotals	**41.84**	**41.29**	**41.02**	**40.86**
Totals	**100**	**100**	**100**	**100**

Table 14. Comparative text and lexical frequencies in GB and GA at 4000-word level.

GENERAL BRITISH (GB)					**GENERAL AMERICAN (GA)**				
IPA	**Lexical**		**Text**		**Text**		**Lexical**		**IPA**
Consonants	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**Consonants**
/n/	7.50	3	7.38	3	7.39	2	7.47	3	**/n/**
/t/	7.42	4	7.60	2	7.65	1	7.81	1	**/t/**
/s/	6.10	5	4.41	5	4.42	5	6.09	5	**/s/**
/l/	5.54	6	3.91	7	3.92	7	5.53	6	**/l/**
/k/	4.97	7	3.29	11	3.30	11	4.96	7	**/k/**
/r/	4.80	8	2.87	13	2.88	13=	4.80	8	**/r/**
/p/	3.49	9	2.15	19	2.16	19	3.48	9	**/p/**
/d/	3.40	10	3.61	9	3.65	9	3.42	10	**/d/**
/m/	3.11	13	2.88	12	2.88	13=	3.11	13	**/m/**
/f/	1.98	15	1.89	21=	1.90	22	1.98	17	**/f/**
/b/	1.67	18	2.71	14	2.72	15	1.68	20	**/b/**
/v/	1.65	19	2.44	17	2.44	17	1.65	21	**/v/**
/ʃ/	1.59	20	0.90	30	0.90	31	1.58	22	**/ʃ/**
/j/	1.30	23	1.22	28	1.19	28	1.23	24	**/j/**
/w/	1.21	25	1.98	20	1.98	21	1.20	26	**/w/**
/z/	1.08	27	1.10	29	1.10	29	1.07	28	**/z/**
/g/	0.99	28	0.88	31	0.88	32	0.99	29	**/g/**
/ʤ/	0.95	30	0.58	36	0.54	36	0.91	30	**/ʤ/**
/ŋ/	0.83	31	0.60	34=	0.60	34	0.84	31	**/ŋ/**
/ʧ/	0.74	33	0.60	34=	0.56	35	0.70	32=	**/ʧ/**
/h/	0.69	35	1.72	24	1.72	25	0.68	34	**/h/**

Table 14. *Cont.*

GENERAL BRITISH (GB)					GENERAL AMERICAN (GA)				
IPA	**Lexical**		**Text**		**Text**		**Lexical**		**IPA**
Consonants	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**Consonants**
/θ/	0.40	37	0.45	38	0.45	37=	0.40	36	**/θ/**
/ð/	0.29	39	3.86	8	3.86	8	0.29	39	**/ð/**
/ʒ/	0.12	42	0.05	44	0.05	44	0.12	42	**/ʒ/**
Subtotals	**61.82**		**59.08**		**59.14**		**61.99**		**Subtotals**
Vowels	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**Vowels**
/ə/	9.23	1	8.50	1	7.30	3	7.52	2	**/ə/**
/ɪ/	7.75	2	6.27	4	6.08	4	7.45	4	**/ɪ/**
/iː/	3.20	11	4.23	6	4.23	6	3.19	11	**/iː/**
/e/	3.18	12	2.32	18	2.32	18	3.16	12	**/ɛ/**
/eɪ/	2.12	14	1.80	23	1.79	24	2.09	16	**/eɪ/**
/æ/	1.86	16	3.38	10	3.62	10	2.17	15	**/æ/**
/aɪ/	1.74	17	1.89	21=	1.86	23	1.71	19	**/aɪ/**
/ɒ/	1.54	21	2.56	15	2.98	12	1.84	18	**/ɑː/**
/ʌ/	1.31	22	1.60	25	1.60	26	1.30	23	**/ʌ/**
/əʊ/	1.22	24	1.28	27	1.28	27	1.21	25	**/oʊ/**
/uː/	1.15	26	2.53	16	2.53	16	1.13	27	**/uː/**
/ɔː/	0.98	29	1.45	26	1.03	30	0.70	32=	**/oɚ/**
/ɑː/	0.78	32	0.57	37	0.26	40	0.36	38	**/ɑɚ/**
/ɜː/	0.71	34	0.64	32	2.10	20	2.76	14	**/ɚ/**
/aʊ/	0.43	36	0.62	33	0.63	33	0.43	35	**/aʊ/**
/eə/	0.32	38	0.42	40	0.45	37=	0.39	37	**/ɛɚ/**
/ʊ/	0.24	40	0.43	39	0.40	39	0.18	41	**/ʊ/**
/ɪə/	0.22	41	0.25	41	0.25	41	0.22	40	**/ɪɚ/**
/ɔɪ/	0.11	43	0.09	42	0.09	42	0.11	43	**/oɪ/**
/ʊə/	0.09	44	0.06	43	0.06	43	0.09	44	**/ʊɚ/**
Subtotals	**38.18**		**40.89**		**40.86**		**38.01**		**Subtotals**
TOTALS	**100.00**		**99.97**		**100.00**		**100.00**		**TOTALS**

Note: "=" indicates that there are two or more phonemes with an equal ranking.

The largest discrepancy in frequency ranks is, as predicted, shown by /ð/, which is at 8 by text frequency in both accents, and at 39 in both by lexical frequency. Otherwise, the consonants are pretty well aligned.

The same cannot be said of the vowels; as Brooks [59] says, "Differences in vowel phonemes constitute the differences between the GB and GA accents" (p. 12). The most surprising differences are at the very top of the table: while the schwa vowel /ə/ has top rank in both text and lexical frequency in GB, it is 2nd in lexical and 3rd in text frequency in GA, and even in GB the percentages are somewhat lower than in the earlier analyses shown in **Table 7** and the 'guesstimate' in **Table 8**. So, Brooks [5] overstated the case when saying: "[/ə/] is the main frequent phoneme of all in spoken English, in every accent … In GB, for example, it constitutes about 10% of running speech" (p. 17). Part of the explanation for the lower ranks of plain /ə/ in GA is that a great many word-final schwas in GA are retroflex /ɚ/ instead.

Among other vowels, the percentages (though not the ranks) for /æ/ show it is, as predicted, more frequent in GA than in GB, because of words like *path, glass*, which in GB have /ɑː/. Despite this, and despite the existence of words such as *father, spa*, which have plain /ɑː/ in both accents, /ɑː/ is much less frequent in GB than in GA. This is because many words in which GB has /ɒ/ have /ɑː/ in GA, and words with <ar> pronounced /ɑː/ in GB have retroflex /ɑɚ/ in GA. Words with /ɔː/ in GB have mostly split between /oɚ/ (if the spelling has <r>) and /ɑː/ otherwise, the latter again contributing to the much higher frequency of /ɑː/ in GA.

Going Just Beyond 44 Phonemes

Up to this point, all the analyses in Part Two have been based strictly on the 44-phoneme sets stipulated in **Tables 2** and **3**. In this section, we offer a pedagogically useful extension to the 44-phoneme sets by introducing analyses which separate out the quasi-phoneme /juː/. In this respect we are following the logic set out in Brooks[5]: "The 2-phoneme grapheme spelling /juː/ (the sound of the whole words *ewe, yew, you* and the name of the letter <u>) – is so frequent that I have infringed my otherwise strictly phonemic analysis to accord the 2-phoneme sequence /juː/ special status as a quasi-phoneme that is important enough to have its own entry as does Carney (1994: 200–202)" (p. 8).

The major reason for according /juː/ quasi-phoneme status is that the letter name vowels /eɪ iː aɪ əʊ juː/, plus /uː/, have interesting properties as a set. First, there is a strong tendency for the letter name vowel plus /uː/ in non-final syllables of polysyllabic words to be spelt with their name letters <a e i o u>. Secondly, there is a strong tendency for those same letter name vowels to be spelt with the corresponding split digraphs both in the final syllables of poly-syllabic words, and in mono-syllables, except that / iː/ in mono-syllables is mainly spelt with other graphemes, and not with <e.e>. Treating /juː/ as a phoneme is pedagogically useful. For full details, see sections 5.7.2, 6.2, 6.3 and 10.17 in Brooks[5].

The results of the 45-item analysis are shown in **Table 15**.

Table 15. Comparative text and lexical frequencies in GB and GA at 4000-word level with the quasi-phoneme/juː/.

GENERAL BRITISH (GB)					**GENERAL AMERICAN (GA)**				
IPA	**Lexical**		**Text**		**Text**		**Lexical**		**IPA**
Consonants	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**Consonants**
/n/	7.54	3	7.39	3	7.39	2	7.50	3	**/n/**
/t/	7.45	4	7.60	2	7.65	1	7.84	1	**/t/**
/s/	6.13	5	4.41	5	4.42	5	6.12	5	**/s/**
/l/	5.56	6	3.92	7	3.92	7	5.55	6	**/l/**
/k/	4.99	7	3.29	11	3.30	11	4.98	7	**/k/**
/r/	4.83	8	2.87	13	2.88	13=	4.82	8	**/r/**
/p/	3.50	9	2.15	19	2.16	19	3.49	9	**/p/**
/d/	3.42	10	3.61	9	3.65	9	3.43	10	**/d/**
/m/	3.13	13	2.88	12	2.88	13=	3.12	13	**/m/**
/f/	1.99	15	1.90	21	1.90	22	1.98	17	**/f/**
/b/	1.68	18	2.71	14	2.72	15	1.68	20	**/b/**
/v/	1.66	19	2.44	16	2.44	16	1.65	21	**/v/**
/ʃ/	1.60	20	0.90	29=	0.90	30=	1.59	22	**/ʃ/**
/w/	1.21	24	1.98	20	1.98	21	1.21	24=	**/w/**
/z/	1.08	25	1.10	28	1.10	28	1.08	26	**/z/**
/g/	0.99	26=	0.88	31	0.89	32	0.99	27	**/g/**
/ʤ/	0.95	28	0.58	37	0.54	37	0.91	28	**/ʤ/**
/j/	0.87	29	0.90	29=	0.90	30=	0.87	29	**/j/**
/ŋ/	0.84	30	0.60	36	0.60	34	0.84	30	**/ŋ/**
/ʧ/	0.75	32	0.61	34=	0.56	36	0.70	32=	**/ʧ/**
/h/	0.69	35	1.72	24	1.72	25	0.69	34	**/h/**
/θ/	0.40	38	0.45	39	0.45	38=	0.40	36	**/θ/**
/ð/	0.29	40	3.86	8	3.86	8	0.29	40	**/ð/**
/ʒ/	0.12	43	0.05	45	0.05	45	0.12	43	**/ʒ/**
Subtotals	**61.67**		**58.8**		**58.86**		**61.85**		**Subtotals**
Vowels	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**%**	**Rank**	**Vowels**
/ə/	9.27	1	8.50	1	7.30	3	7.55	2	**/ə/**
/ɪ/	7.78	2	6.27	4	6.08	4	7.48	4	**/ɪ/**
/iː/	3.22	11	4.23	6	4.23	6	3.20	11	**/iː/**
/e/	3.20	12	2.32	17	2.32	17	3.17	12	**/ɛ/**
/eɪ/	2.13	14	1.80	23	1.79	24	2.09	16	**/eɪ/**
/æ/	1.87	16	3.38	10	3.62	10	2.18	15	**/æ/**
/aɪ/	1.75	17	1.89	22	1.86	23	1.72	19	**/aɪ/**

Table 15. *Cont.*

Vowels	%	Rank	%	Rank	%	Rank	%	Rank	Vowels
/ɒ/	1.54	21	2.57	15	2.98	12	1.85	18	**/ɑː/**
/ʌ/	1.31	22	1.60	25	1.60	26	1.31	23	**/ʌ/**
/əʊ/	1.22	23	1.28	27	1.28	27	1.21	24=	**/oʊ/**
/ɔː/	0.99	26=	1.45	26	1.03	29	0.70	32=	**/oɚ/**
/ɑː/	0.78	31	0.57	38	0.26	41	0.36	39	**/ɑɚ/**
/ɜː/	0.71	33=	0.64	32	2.10	20	2.77	14	**/ɚ/**
/uː/	0.71	33=	2.21	18	2.24	18	0.76	31	**/uː/**
/juː/	0.45	36	0.61	34=	0.57	35	0.37	38	**/juː/**
/aʊ/	0.43	37	0.63	33	0.63	33	0.43	35	**/aʊ/**
/eə/	0.32	39	0.42	41	0.45	38=	0.39	37	**/ɛɚ/**
/ʊ/	0.24	41	0.43	40	0.40	40	0.18	42	**/ʊ/**
/ɪə/	0.23	42	0.25	42	0.25	42	0.22	41	**/ɪɚ/**
/ɔɪ/	0.11	44	0.09	43	0.09	43	0.11	44	**/oɪ/**
/ʊə/	0.09	45	0.06	44	0.06	44	0.09	45	**/ʊɚ/**
Subtotals	**38.35**		**41.20**		**41.14**		**38.14**		**Subtotals**
TOTALS	**100.02**		**100.00**		**100.00**		**99.99**		**TOTALS**

Note: "=" indicates that there are two or more phonemes with an equal ranking.

4. Conclusions

Of all the previous analyses mentioned in Part One, only seven stood up to rigorous analysis against our 44-phoneme sets in both accents. Of the four previous analyses relevant to GB/RP, even the most recent is 40 years old, though it did provide the basis for the only useful overtime comparisons reported in this article; see Section 1.7. Of the three previous analyses relevant to US accents, none, in our opinion, provided data on the GA accent as such. They did provide useful information on historical analyses of various US accents, though even the most recent is now 60 years old. We therefore offer our analyses of the lexical frequencies and text frequencies of phonemes in both accents as interesting in themselves, and as the bases for future overtime comparisons. The largest database cited in the seven previous analyses we have analysed is that of Dewey [18]: 100,000 words. Yet even that is orders of magnitude smaller than the billion+ words in the COCA database, which is therefore a much more secure basis for current and future analyses.

The finalised results of our 44-phoneme analysis in both accents are shown in **Table 14**. We intend the 45-item analysis shown in **Table 15** to be a pedagogically useful outcome of all this work. Cochrane and Brooks [53] traced the influence of the 'satpin assumption' on many initial reading schemes in Britain, the assumption being that the graphemes <satpin> and their most frequent correspondences with phonemes, /s æ t p ɪ n/ and near variants of it, offer an optimal starter set for phonics schemes for beginning readers. But it must be remembered that the origins of that assumption spring from a structured literacy programme based upon lexical frequencies in an unspecified US accent analysed by Hanna et al. in a work published in 1966. Cochrane & Brooks (in preparation) will demonstrate an alternative starter set to *satpin* based on up-to-date data on the GB accent produced in this article, and specifically using not lexical but text frequencies, since these are what writers produce and readers encounter.

Implication of Our Study

Key implications of the data presented in this article are:

1. The data fill critical accent-inventory gaps. Ours is the first fully phonetic count of the GA accent. Previous studies lack both modern transcription rigour and representative sampling. Our study provides parallel analyses of the GA and GB accents. By applying the same 44-(45)-phoneme set, corpus size, and pipeline to both accents, we create directly comparable benchmarks. This symmetry is unprecedented and is essential for future rigorous, cross-dialectal work. The GB aspect of the study updates the GB inventory after nearly 40 years. The COCA-based data capture real shifts in vowel qualities and prosodic patterns in the GB accent, ensuring that the study's GB rank orders reflect today's usage rather than archival speech.

2. The data provide methodological innovations with immediate utility:
    a. Balanced, multi-register sampling overcomes narrow-sample biases of many previous studies, giving robust counts for everyday and formal speech in both the GB and GA accents.
    b. Demonstrating that frequencies stabilize only after the top 3000- to 4000-word types in the corpus provides concrete guidance on corpus size requirements for reliable phoneme models. This advice is critical for any research or tool relying on phoneme probability data.
    c. By treating /juː/ as a quasi-phoneme, we reduce distributional bimodality and better reflect learner input. This refinement immediately benefits both probabilistic modelling and instructional sequencing.
3. The data set offers potential refinements to speech communication research:
    a. Accurate phoneme frequencies and rank orders underpin phonotactic-probability calculations, which predict the likelihood of phoneme sequences in perception, lexical access and serial recall. Our comprehensive GB and GA accent tables allow those models to reflect current distributions, improving fit for phenomena such as non-word acceptability and phoneme surprisal in continuous speech.
    b. Our frequency database serves as the empirical basis for mapping phoneme ranks to grapheme sequences in both accents. This supports the creation of more precise phonotactic methods for predicting spelling patterns, error rates in reading-aloud tasks and the design of literacy interventions aligned to accent-specific frequency profiles.
    c. Emerging speech-communication technologies (feedback systems for pronunciation training, articulatory-feedback apps, adaptive listening tools, etc.) depend on realistic phoneme rank orders. Our standardized GB–GA counts provide the benchmarks needed to calibrate those systems for target-accent modelling and to evaluate learner progress against authentic usage.
4. The data informs the design of future phonics and literacy intervention schemes. As already mentioned, many current schemes rely on outdated, US-based, lexical frequency rank lists. Introducing GB and GA phoneme-introduction orders that mirror modern usage within the corpus should speed early decoding, reduce learner confusion and improve reading-aloud accuracy in both accents.
5. Text-to-speech and automatic speech recognition technology can be enhanced by our frequency database. Their back-ends use phoneme n-gram probabilities derived from frequency tables. Incorporating our comprehensive, up-to-date GB and GA accent data will reduce mispredictions and lead to more natural synthesis and higher recognition accuracy.

In summary, up-to-date, standardised phoneme frequencies for the GB and GA accents have the potential to enable advances in theoretical phonetics, psycholinguistics, speech technology, phonotactics, and reading and spelling pedagogy. We provide a fully documented, reproducible pipeline, from corpus filtering through phonemic mapping, with all code and tables provided as open resources (**Supplementary Materials**).

Supplementary Materials

The supporting information can be downloaded at: https://journals.bilpubgroup.com/files/FLS-10858-Supplementary-Material.xlsx.

Author Contributions

G.B. wrote about 85% of the text, devised the framework for all 15 tables and provided the whole contents of the first, fourth to ninth tables, including the re-analysis of data from Whitney, Dewey, Trnka and Hanna et al. G.B. and B.B. agreed on the 44-phoneme inventories for GB and GA, and set them out in the second and third tables; and checked the transcriptions of the first 4000 words in GB and GA, respectively. A.G.I. accessed the COCA database, implemented the transcriptions of the first 20,200 words in both accents, undertook the calculation of frequencies of the phonemes in both accents, which resulted in the tenth to fifteenth tables

and provided extra text about the background of phonetic analysis of the GA accent and the implications of the study. G.C. provided the analysis underlying the section on changes over time in the GB accent and wrote that section. G.C. contributed to the abstract and conclusions sections and acted as desk editor of the entire article and as the submitting author. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable to this study.

Data Availability Statement

The dataset supporting the reported results in our study can be found in Supplementary Materials.

Acknowledgements

The authors wish to express their deep gratitude to the following people. To Dr Marie Ernestová, former Senior Lecturer in the English Department of the Faculty of Education, University of South Bohemia, Czech Republic (and once Bohumil Trnka's student) and to her colleague Marta Němcová, BA, of the English and American Studies and Romance Studies Library of Charles University, Faculty of Arts, Prague, for chasing up publication details of Trnka (1935). To Prof. P. David Pearson and Prof. Jane Setter for helpful null returns about recent phoneme frequency data in the GA accent; to Oliver Allchin of the University of Sheffield library for providing literature searches; to Robert Thresh, excellent amanuensis. To Devakanni Subramanian from Spell Bee International, who helped the first author get the license to have the COCA Corpus for the analysis. To Christopher Brooks, for cross-checking some statistical matters. To Imogen Wilkinson, for providing some crucial IT support.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Appendix A. Search Strategy

(TITLE-ABS-KEY (phoneme* OR phonetic* OR phonolog*) AND TITLE-ABS-KEY (british OR britain OR uk OR "received pronunciation") AND TITLE-ABS-KEY (america* OR "united states" OR us) AND TITLE-ABS-KEY (english))

(TITLE-ABS-KEY (phoneme* OR phonetic* OR phonolog*) AND TITLE-ABS-KEY (british OR britain OR uk OR "received pronunciation" OR "general british") AND TITLE-ABS-KEY ("general american" OR ga OR america* OR "united states" OR us) AND TITLE-ABS-KEY (english))

196 results scopus

—

phoneme* OR phonetic* OR phonolog*

AND

"general british" OR GB OR british OR britain OR "received pronunciation" OR RP OR UK OR "united Kingdom"

AND

"general american" OR GA OR america* OR "united states" OR US OR USA

AND

English

228 results Scopus

—

phoneme* OR phonetic* OR phonolog*

AND

"general british" OR GB OR british OR britain OR "received pronunciation" OR RP OR UK OR "united Kingdom"

AND

"general american" OR GA OR america* OR "united states" OR US OR USA OR "standard american"

341 results Scopus

—

phoneme* OR phonetic* OR phonolog*

AND

"general british" OR GB OR british OR britain OR "received pronunciation" OR RP OR UK OR "united kingdom"

AND

"general american" OR GA OR america* OR "united states" OR US OR USA

AND

English

34 results WoS

—

phoneme* OR phonetic* OR phonolog*

AND

"general british" OR GB OR british OR britain OR "received pronunciation" OR RP OR UK OR "united kingdom"

AND

"general american" OR GA OR america* OR "united states" OR US OR USA

56 results WoS

—

References

[1] Knowles, G., 1987. Patterns of Spoken English: An Introduction to English Phonetics. Longman: Harlow, UK.

[2] Coxe, M.S., 1937. A history of the spelling of English phonemes [PhD thesis]. Louisiana State University: Baton Rouge, LA, USA. Available from: https://repository.lsu.edu/gradschool_disstheses/7861/ (cited 2 March 2025).

[3] Carney, E., 1994. A Survey of English Spelling. Routledge: London, UK.

[4] Gontijo, P.F., Gontijo, I., Shillcock R., 2003. Grapheme-phoneme probabilities in British English. Behavior Research Methods, Instruments and Computers. 35(1), 136–157. DOI: http://dx.doi.org/10.3758/bf03195506

[5] Brooks, G., 2015. Dictionary of the British English Spelling System. Open Book Publishers: Cambridge, UK. Available from: http://www.openbookpublishers.com/reader/325 (cited 2 March 2025).

[6] Gimson, A.C., Cruttenden, A., 2014. Gimson's Pronunciation of English, 8th ed. Edward Arnold: London, UK.

[7] Hanna, P.R., Hanna, J.S., Hodges, R.E., et al., 1966. Phoneme-grapheme Correspondences As Cues to Spelling Improvement. US Department of Health, Education and Welfare: Office of Education, US Government Printing Office: Washington, DC, USA. Available from: http://files.eric.ed.gov/fulltext/ED128835.pdf (cited 2 March 2025).

[8] Gimson, A., 1970. An Introduction to the Pronunciation of English, 2nd ed. Edward Arnold: London, UK.

[9] Fry, D.B., 1947. The frequency of occurrence of speech sounds in southern English. Archives Néerlandaises de Phonétique Expérimentale. 20, 103–106.

[10] Denes, P.B., 1963. On the statistics of spoken English. Journal of the Acoustical Society of America. 35, 892–904. DOI: https://doi.org/10.1121/1.1918622

[11] Denes, P.B., 1964. On the statistics of spoken English. Zeitschrift für Phonetik, Sprachwis-senschaft und Kommunikationsforschung. 17(1–6), 51–72. DOI: https://doi.org/10.1524/stuf.1964.17.16.51

[12] Merriam-Webster, Inc. Staff, 2008. Merriam-Webster's Advanced Learner's English Dictionary. Merriam-Webster Inc.: Springfield, MS, USA.

[13] Merriam-Webster Inc., 2003. Merriam-Webster's Collegiate Dictionary, 11th ed. Merriam-Webster Inc.: Springfield, MS, USA.

[14] Johnson, K., Ladefoged, P., 2014. A course in phonetics, 7th ed. Cengage Learning. Inc.: Boston, MA, USA.

[15] Mines, M.A., Hanson, B.F., Shoup, J.E., 1978. Frequency of occurrence of phonemes in conversational English. Language and Speech. 21(3), 221–241. DOI: https://doi.org/10.1177/002383097802100302

[16] Whitney, W.D., 1874. The elements of English Pronunciation. In Oriental and Linguistic Studies, Second Series: The East and West; Religion and Mythology; Orthography and Phonology; Hindu Astronomy. Scribner: New York, NY, USA. pp. 202–276. Available from: https://archive.org/details/orientallinguist02whituoft/page/272/mode/2up (cited 2 March 2025).

[17] Whitney, W.D., 1874. The proportional elements of English utterance. Available from: https://www.jstor.org/stable/pdf/2935822.pdf

[18] Dewey, G., 1923. Relative Frequency of English Speech Sounds: Harvard Studies in Education 4. Harvard University Press: Cambridge, MA, USA. Available from: https://archive.org/details/in.ernet.dli.2015.18294/page/n1/mode/2up (cited 2 March 2025).

[19] Kenyon, J.S., Knott, T.A., 1953. A Pronouncing Dictionary of American English. Merriam-Webster Inc.: Springfield, MS, USA.

[20] Carley, P., Mees, I., 2020. American English Phonetics and Pronunciation Practice. Routledge Taylor & Francis Group: New York, NY, USA.

[21] Fowler, M., 1957. Herdan's statistical parameter and the frequency of English phonemes. In: Studies Presented to Joshua Whatmough, 45. De Gruyter: The Hague, The Netherlands. pp. 47–52.

[22] Wang, W.S.-Y., Crawford, J., 1960. Frequency studies of English consonants. Language and Speech. 3(3), 131–139. DOI: https://doi.org/10.1177/002383096000300302

[23] Gerber, S.E., Vertin, S., 1969. Comparative frequency counts of English phonemes. Phonetica. 19, 133–141.

DOI: https://doi.org/10.1159/000258622

[24] Trubetzkoy, N.S., 1969. Principles of Phonology. Baltaxe, C.A. (Trans.). University of California Press: Berkeley, CA, USA.

[25] Trnka, B., 1935. A Phonological Analysis of Present-Day Standard English (Studies in English by Members of the English Seminar of Charles University. 5, 1–175). Faculty of Philosophy, Charles University: Prague, Czech Republic. (in German)

[26] Berndt, R.S., Reggia, J.A., Mitchum, C.C., 1987. Empirically derived probabilities for grapheme-to-phoneme correspondences in English. Behavior Research Methods, Instruments, and Computers. 19, 1–9. DOI: https://doi.org/10.3758/BF03207663

[27] Atkins, R.E., 1926. An analysis of the phonetic elements in a basal reading vocabulary. The Elementary School Journal. 26(8), 596–606. Available from: http://www.jstor.org/stable/994876

[28] Hayden, R.E.,1950. The relative frequency of phonemes in general American English. Word. 6(3), 217–223. DOI: https://doi.org/10.1080/00437956.1950.11659381

[29] Thorndike, E.L., 1921. The Teacher's Word Book. Bureau of Publications, Teachers College, Columbia University: New York, NY, USA.

[30] Carroll, J.B., 1952. Transitional Probabilities of English Phonemes. Available from: https://files.eric.ed.gov/fulltext/ED022182.pdf?utm_source (cited 12 April 2025).

[31] Wang, W.S.-Y., 1965. Reviewed Work: Tables of Transitional Frequencies of English Phonemes by Joseph H. D. Allen, Jr., Lee S. Hultzén, Murray S. Miron. 4(3), 525–529. DOI: https://doi.org/10.2307/411797

[32] French, N.R., Carter, C.W., Koenig, W., 1930. The words and sounds of telephone conversations. Bell Systems Technical Journal. 9, 290–324. DOI: https://doi.org/10.1002/j.1538-7305.1930.tb00368.x

[33] French, N.R., Koenig Jr., W., 1929. The frequency of occurrence of speech sounds in spoken English. Journal of the Acoustical Society of America. 1(34), 110–120. DOI: https://doi.org/10.1121/1.1901472

[34] Voelker, C.H., 1937. A comparative study of investigations of phonetic dispersion in connected American speech. Archives Néerlandaises de Phonétique Expérimentale. 13, 138–157.

[35] Tobias, J.V., 1959. Relative occurrence of phonemes in American English. Journal of the Acoustical Society of America. 31, 631. DOI: https://doi.org/10.1121/1.1907766

[36] Irwin, O.C., 1947. Infant speech: consonantal sounds according to place of articulation. Journal of Speech and Hearing Disorders. 12(4), 397–404. DOI: https://doi.org/10.1044/jshd.1204.397

[37] Irwin, O.C., 1948. Infant speech: Development of vowel sounds. Journal of Speech and Hearing disorders. 13(1), 31–34. DOI: https://doi.org/10.1044/jshd.1301.31

[38] Delattre, P., 1965. Comparing Phonetic Features of English, French, German and Spanish. Groos:Heidelberg, Germany. Available from: https://archive.org/details/comparingphoneti0000unse (cited 12 April 2025).

[39] Isengel'dina, A.A., 1975. Some questions of phonological statistics. Linguistics. 13(146), 15–30. DOI: https://doi.org/10.1515/ling.1975.13.146.15

[40] Carterette, E., Jones, M., 1974. Informal Speech: Alphabetic and Phonemic Texts with Statistical Analyses and Tables. University of California Press: Berkeley, MA, USA.

[41] The Language Nerds, n.d. Most common sounds in spoken English. Available from: https://thelanguagenerds.com/2019/most-common-sounds-in-spoken-english/ (cited 12 April 2025).

[42] Zipf, G.K., 1929. Relative frequency as a determinant of phonetic change. Harvard studies in classical philology. 40, 1–95. DOI: https://doi.org/10.2307/310585

[43] Jones, D., 1912. Phonetic Readings in English, 1st ed. G.E. Stechert & Co.: New York, NY, USA. Available from: https://archive.org/details/phoneticreadings00jone (cited 12 April 2025).

[44] Scott, N.C., 1942. English Conversations in Simplified Phonetic Transcription. Heffer: Cambridge, UK.

[45] Ramsaran, S., 1990. RP: Fact and fiction. In: Ramsaran, S. (Ed.). Studies in the Pronunciation of English: A Commemorative Volume in Honour of A.C. Gimson. Routledge: London, UK. pp. 178–190.

[46] Wells, J.C., 2010. John Wells's phonetic blog: believing descriptions. Available from: https://phonetic-blog.blogspot.com/2010/11/believing-descriptions.html (cited 27 June 2025).

[47] Wells, J.C., 2014. Sounds Interesting. Cambridge University Press: Cambridge, UK.

[48] Windsor Lewis, J., 1990. HappY land reconnoitred: the unstressed word-final-y vowel in General British pronunciation. In: Ramsaran, S. (Ed.). Studies in the Pronunciation of English: A Commemorative Volume in Honour of A. C. Gimson. Routledge: London, UK. pp. 159–167.

[49] Roberts, A.H., 1965. A Statistical Linguistic Analysis of American English, Doors of Language Series Practice 8. Mouton: The Hague, The Netherlands.

[50] Trnka, B., 1966. A Phonological Analysis of Present-Day Standard English: Alabama Linguistic and Philological Series 17. University of Alabama: Tuscaloosa, AL, USA.

[51] Dušková, L., 2014. English studies at Charles University and the Prague Linguistic Circle: The contribution of English Studies to the Circle's constitution and linguistic theories. La Linguistique. 50(1), 93–118. DOI: https://doi.org/10.3917/ling.501.0093

[52] Fowler, F., Fowler, H., 1928. The Pocket Oxford Dic-

tionary of Current English. Clarendon Press: Oxford, UK.

[53] Cochrane, G., Brooks, G., 2022. Where should phonics teaching start? 'satpin' and its origins, rivals and implications. British Educational Research Journal. 48(6), 1198–1215. DOI: https://doi.org/10.1002/berj.3822

[54] Thorndike, E.L., Lorge, I., 1944. The Teacher's Word Book of 30,000 Words. Bureau of Publications, Teachers College, Columbia University: New York, NY, USA.

[55] Davies, M., 2020. Corpus of Contemporary American English (COCA). Available from: https://repository.gonzaga.edu/corp_coca/27/ (cited 27 June 2025).

[56] Davies, M., 2010. The Corpus of Contemporary American English as the first reliable monitoring corpus of English. Literary and Linguistic Computing. 25(4), 447–464. DOI: https://doi.org/10.1093/llc/fqq018

[57] Dockum, R., Bowern, C., 2019. Swadesh lists are not long enough: drawing phonological generalizations from limited data. Language Documentation and Description. 16, 35–54. DOI: https://doi.org/10.25894/ldd112

[58] Bowern, C., 2024. Sample sizes for Australian phonetics corpora. Available from: https://campuspress.yale.edu/clairebowern/sample-sizes-for-australian-phonetics-corpora/ (cited 27 June 2025).

[59] Brooks, G., 2021. The linguistic base of initial reading and spelling in English: a tutorial review. Education 3–13. 49(1), 10–28. DOI: https://doi.org/10.1080/03004279.2020.1824699