2013

Identifying Authors by Phonoprints in Their Characters' Names: An Exploratory Study

Wendy Baker-Smemoe
Brigham Young University, wendy_baker@byu.edu

Brad Wilcox
Brigham Young University

Bruce L. Brown
Brigham Young University

Sharon Blake
Brigham Young University

Justin Bray
Brigham Young University

Original Publication Citation

Wilcox, B., Brown, B. L., Baker-Smemoe, W., Black, S., & Bary, J. (2013). Identifying authors by phonoprints in the characters' names: An exploratory study. Names, 61, 104-125.

NAMES, Vol. 61 No. 2, June 2013, 101–21

Identifying Authors by Phonoprints in Their Characters' Names: An Exploratory Study

Brad Wilcox, Bruce L. Brown, Wendy Baker Smemoe, Sharon Black, and Justin Bray
Brigham Young University, USA

If authors put words together in ways that can be recognized as wordprints (Hilton, 1990; Morton, 1979; Archer et al., 1997), do they put sounds together in identifiable ways when they invent names? Could they have unique sound prints (phonoprints) as well? This exploratory study compared phonemic patterns of fictional names in the poorly written *Manuscript Story* by Spalding and the extremely well-written *Lord of the Rings* and related works by J. R. R. Tolkein with names from an authentic public record, the nineteenth-century US Census. Phonotactic probabilities were determined using a calculator (Vitevitch and Luce, 2004) available on the Internet. When multivariate patterns of mean phonotactic probabilities at each ordinal phoneme position were considered, phonoprints emerged that merit further examination.

KEYWORDS author identification, phonoprints, fictional character names, authentic names, nineteenth-century census, J. R. R. Tolkein, phonotactic probabilities

An elementary school teacher noticed that children reading the Harry Potter fantasy series were writing their own fantasies, with imaginative characters, fantastic settings, and strange names. Some students shortened or expanded familiar names; others compounded two known words to make a new name. Each child seemed to be settling into his or her recognizable pattern.

Fiction authors are more sophisticated than children, but do they also have their own individual patterns and preferences when they invent names? Do these patterns differ from those of other authors or those found in authentic names?

Some linguists believe that each author creates an individual and identifiable *word-print* by the way he or she puts words together (Hilton, 1990; Morton, 1979; Archer et al., 1997). Although they are not as unique as fingerprints and are sometimes

 DOI 10.1179/0027773813Z.00000000048

tentative and difficult to define (Croft, 1981), wordprints can be used to determine probability of a writer's identity and are used regularly in verifying document authorship (Baily, 1979; Holmes, 1994; Zheng et al., 2003; Iqbal et al., 2010). If authors put words together in ways that can be recognized, might they also — consciously or subconsciously — put sounds together in consistent individual patterns when they invent words? Could each have his or her own sound print (*phonoprint*) as well?

Vitevitch et al. (1997) reported a significant correlation between what they called phonotactic probability — the chance of certain sounds coming in certain positions in English — and what they labeled *neighborhood density* — sequences of sounds commonly found in close proximity. They found that consonant-vowel-consonant content words comprised of common sounds tend to have many predictable lexical neighbors. For example, some of the "neighbors" for the word *cap* would be *cat*, *can*, *cost*, *and tap*. Another example which comes from historical Germanic and Latin roots would be that /k/ is often followed by /w/ as in *queen* or *quick*. When authors create fictional names, do they unwittingly create them from similar phonotactic neighborhoods? Do they group sounds in typical ways or pair non-typical sounds consistently? Regardless of varied backgrounds and skill levels, do authors of fictional names use phonemic patterns that differ from those seen in nonfiction names? This study was undertaken to explore such possibilities.

Purpose

Whether names are authentic, adapted, or created, processes of generating them have been studied for years (Francis, 1981). However, few have examined the process author by author. The purpose of this study was to analyze and compare the phonemic patterns of the male names found in a little-known manuscript by Solomon Spalding (considered by many to be poor fiction), and *The Lord of the Rings* and related works by J. R. R. Tolkein (considered by many to be excellent fiction) with authentic male names from a public source: the US Census records of the nineteenth century, the time at which Spalding produced his manuscript. Male single-word names were used because of their prevalence in the selected texts. Recognizing the broad diachronic comparisons between works attributed to the nineteenth and twentieth centuries, we have considered this study as merely exploration to determine if differences between authors and between fictional and authentic names merit further investigation.

Phonotactic research

Typical methodologies for name study include structural analysis and contemporary or historical comparison — allowing researchers to determine whether a word is part of a specific language (Wu, 2010). Some researchers ask native speakers to confirm whether a word "sounds" like the language or not (Young, 2004); others use a corpus of a particular language for historical linguistic analysis (Downey et al., 2008).

None of these methods was completely adequate for studying the names in our selected works. Traditional structural analysis of the names would require data not currently available. Comparing the names to words in other languages would be

tentative, since both novelists invented the languages from which character names supposedly came; thus native speakers would not exist.

This study focused on *phonotactic probability*, previously defined as the general frequency of occurrence of phonological segments and sequences of segments in a given language (Jusczyk et al., 1994). For example, the vowel sounds found at the beginning of the words *eat* and *if* are more common in English than the vowel sounds found at the beginning of *alms* and *oink*. Similarly, consonant sounds such as those found at the beginning of *love*, *kiss*, *ton*, and *new* are all much more common in English than those found at the beginning of *young* and *whip* (Kessler and Treiman, 1997).

All sounds (phonemes) and pairs of sounds (biphonemes, labeled *bifones* by Vitevitch and Luce [2004]) have varying levels of probability in English, depending on their position in any given word or name regardless of origin. For example, *Adam* and *Solomon* both have their origins in Hebrew, yet the average probability that in English all the phonemes and bifones would come in the order they do is higher for Solomon (Phonemes = .0716; bifones = .0081) than for Adam (Phonemes = .0231; bifones = .0020). The probability that in English the sounds in the word *Solomon* would come in the position they do is about 7%, and the probability of the pairs of sounds found in *Solomon* coming in their ordinal positions is less that 1% on average. In contrast, the sounds in *Adam* have a 2% probability of coming in the order they do, with the pairs of sounds having less than two-tenths of 1% probability of occurring. Phonotactic probability does not measure how common a word or name is. There may be more men named Adam than Solomon. Phonotactic probability deals only with prevalence of the sound sequence.

These determinations can be made using the probability calculator developed by Vitevitch and Luce (2004), available on the Internet (<http://www.bncdnet.ku.edu/cgi-bin/DEEC/post_ppc.vi>), which compares each word to the Brown Corpus, a frequency database of standard American English created by Kucera and Francis (1967). Each word can be entered phonemically or phonetically using the computer-readable transcription method called Klattese developed by Dennis Klatt (<http://129.237.66.-221/VLbrmic.pdf>). In this study, words were entered phonemically because little is known about the context-conditioned variations of many of the names on a phonetic level. Klatt uses keys available on a keyboard to represent unique sounds. For example, Edgar was entered as *Edgx*, Erchamion was rendered as *xCamian*, and Borthand was *borT@nd*. The output of the calculator contains the position-specific probability for each phoneme, the position-specific probability for each bifone, and the sum of all phoneme and bifone probabilities.

Phonotactics, the body of rules that determine the constraints on the location of sounds in the syllable structure of languages, has been used extensively to study patterns in English names (Whissell, 2001; Shih, 2012) including the differences between male and female names and ways phonological cues can predict gender (Wright, 2012; Starr, 2012). Phonotactic probabilities have been used in relation to language, memory, speech, and hearing. For example, researchers have examined the prominence of certain beginning and ending sounds in words and syllables, exploring how such word structures influence spoken-word recognition (Vitevitch, 2002; Vitevitch and Luce, 1999). Others have looked at similarities in certain groups of

words (Bailey and Hahn, 2001) relevant to language learning and development (Storkel, 2001; 2003). Studies have examined speech errors and phonotactic constraints (Dell et al., 2000), as well as infants' sensitivity to the sound patterns of native language words (Jusczyk et al., 1993).

Only a few studies have examined invented words, including one demonstrating that infants of similar ages could discriminate which non-words contained sounds more common or less common in their native languages (Jusczyk et al., 1993). Another explored the processing of non-words by deafened adults with cochlear implants (Vitevitch et al., 2002). We know of no other studies of phonotactic probabilities provided by the Phonotactic Calculator having been applied to fictional or authentic name sources. Thus we are using an established method to perform a new analysis.

Three name sources

The three names sources for this study were a manuscript by little-known nineteenth-century author Solomon Spalding, *The Lord of the Rings* and related works by J. R. R. Tolkein, and the US Census records of the nineteenth century. The Spalding manuscript, dismissed by many as poor fiction, was included for its many author-invented names that are not found or used elsewhere. Tolkein's novels, on the other hand, are generally recognized as examples of excellent fiction and were included for their unusual names.

Names from the Spalding manuscript

Around 1800 Solomon Spalding (or Spaulding; 1761–1816) wrote a little-known fictional story titled *Manuscript Story* (also referred to as *Manuscript Lost* or *Manuscript Found*), a fictional account of a lost civilization of Native Americans who built earth mounds in central and eastern United States. A group from Rome, traveling to England around the time of Christ, were blown off course and landed on the North American continent where they interacted with these native people.

Spalding's fictional work includes unusual place names (e.g., *Tolanga*) and groups of people (e.g., *Sciotans* and *Kentucks*). Common personal names like *Tom* are used, but we identified 61 personal names not found elsewhere for use in this study (e.g., *Lobaska*, *Bombal*, *Lamack*, *Helicon*).

Pronunciation of names

Spalding did not provide a pronunciation guide for *Manuscript Story*, so two different lists of pronunciations were used in this study: one based on general decoding guidelines for English (Eldredge, 2005), a second based on decoding guidelines for Latin, since the story's travelers migrated from Rome.

An ANOVA analysis revealed that the English and Latin pronunciations differed significantly in phoneme probability, the English pronunciations being more like standard American English in the Kucera and Francis corpus (1967), both for phonemes (.289 compared to .220, $F[1, 486] = 168.51$, $p < .0001$, $R^2 = .258$) and for bifones (.017 compared to .009, $F[1, 364] = 134.79$, $p < .0001$, $R^2 = .270$). However, considering the patterns of probabilities for each of the four ordinal positions of phonemes and

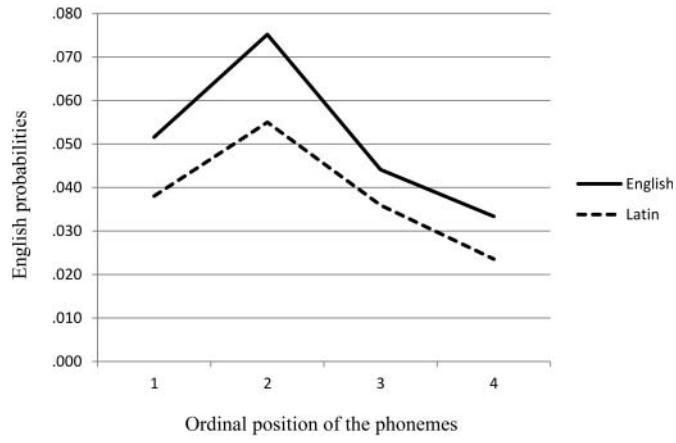

FIGURE 1 A comparison of English pronunciations and Latin pronunciations of the Spalding names in their phoneme probabilities at each of the first four ordinal positions.

bifones showed the two pronunciations virtually identical (see Figure 1); thus we included only English pronunciations in the study.

Ordinal position comparisons are problematic across names of different lengths; thus we used a telescoping design including four analyses: one for all names of four or more phonemes (122 paired English and Latin names, analyzed for four ordinal positions), one for all names of five or more phonemes (106 paired names, analyzed for five ordinal positions), one for all names of six or more phonemes (72 pairs), and one for all names of seven or more phonemes (22 pairs). Too few names had eight phonemes to provide analysis. The results of the three latter analyses were consistent with those from the analysis of all four-or-more-phoneme names.

Phonotactic descriptions

While all 61 personal names in the Spalding manuscript were included in the pronunciation analysis, female names were not used in the remainder of this study; thus we examined 55 male names by Spalding. The Spalding name containing the most phonemes was *Drafolick*, with a total of eight. Three names containing only four phonemes, *Como*, *Droll*, and *Kato*. *Sambol* and *Baska*, were the most like standard English in the Brown Corpus, with average phoneme probabilities of .0712 and .0666. *Ulipoon* was the least like English in the Brown Corpus, with an average phoneme probability of .0330. Considering bifones, the name most like English in the corpus was *Hamul*, with an average bifone probability of .0070. The name least like standard English was *Boakim*, with an average bifone probability of .0010. Overall, the longer the name, the less it was like standard American English.

Names from The Lord of the Rings

The *Lord of the Rings* trilogy, composed of *The Fellowship of the Ring, The Two Towers*, and *The Return of the King*, was first published by J. R. R. Tolkien (1892–1973) between 1954 and 1955. *The Hobbit* was published in 1937. After Tolkien's death a related collection of legends and tales about pre-middle earth was published

in 1977 called *The Silmarillion*. Some claim the entire story reflects Tolkien's interest in Germanic and Celtic mythology and folklore (Chance, 2004). These works introduced civilizations and cultures with different languages and numerous unusual names. For ease of reading in this study, all Tolkien's works set in middle earth were labeled with one title: *The Lord of the Rings*.

Tolkien, a linguist who became a professor of Anglo-Saxon in 1933, was well versed in many languages, including medieval and modern Germanic and Celtic tongues. With his knowledge and love of languages, he created several artificial languages (based mostly on languages like Latin, Welsh, Finnish, and Old Norse); vestiges of these invented languages appear in his fiction.

Character names in *The Lord of the Rings* were derived from several sources. Tolkien explained that many of the names were from invented Elvish languages (e.g., *Legolas*, *Elrond*, and *Galadriel*); others were translated from these languages into English equivalents (e.g., *Treebeard*, *Skinbark*, *Leaflock*). Some were derived from Old English to appear ancient (e.g., *Theoden*, *Eomer*, *Erowyn*). Some seem to have been derived from the author's childhood experiences, such as Sam's surname *Gamgee*, a word used for a cotton mill in the small town of Sarehole where Tolkien grew up.

Pronunciations of names

Pronunciations for names in *The Lord of the Rings* were based on pronunciation transcriptions from audio recordings of Tolkien reading his own books (<http://inogolo.com/guides/lord+of+the+rings>). Pronunciations used in the film trilogy directed by Peter Jackson were also consulted.

Phonotactic descriptions

This study used 197 male names from Tolkien that are not found in other sources, selected from an Internet list (<http://lotr.wikia.com/-wiki/List_of_characters>) and verified using the online Encyclopedia of Arda (<http://www.glyphweb.com/arda/>). As few surnames or titles appear in the text, only first names were included. Names given to two or more characters were used only once in the study.

The name with the most phonemes in *The Lord the Rings* was *Celebrimbor*, with 11. *Oin* had the least, with two phonemes. Among the names that were the most like standard American English in the Brown Corpus were *Sauron*, *Saruman*, and *Beren*, with average probabilities of .0792, .0723, and .0715 respectively. The least like English in the corpus was *Azog*, with an average phoneme probability of .0098. Focusing on the bifones, the name that was most like standard English was *Indis*, followed closely by *Sauron* and *Saruman*, with average bifone probabilities of .0126, .0111, and .0101 respectively. The names least like English in the corpus were *Oin*, *Azog*, and *Olwe*, with average bifone probabilities of .0001, .0002, and .0003. Names that were longer were more like standard American English.

Nineteenth-century names

Throughout the 1800s both male and female given names were gathered and reported on the US Census, but, to be consistent with the prevalence of male single-word names in the books, only the 100 most common male first names on the census were

used (<http://users.erols.com/dgalbi/names/us200.htm>; see also Erickson et al., 2008). Their origins were identified using the *Dictionary of First Names* (Hanks and Hodges, 1995). Although English is a Germanic language and name similarities and overlaps occur, the origins provided in the dictionary were used with no alterations.

Origins of names

In the United States in the nineteenth century, the 10 most popular male names were John (from Hebrew), William (Germanic), James (Hebrew), George (Greek), Charles (Germanic), Thomas (Aramaic), Joseph (Hebrew), Henry (Germanic), Samuel (Hebrew), and Robert (Germanic). Of the 100 most popular male names, 34 were of Hebrew origin, 22 were Germanic, 16 were English, 11 were Latin, 9 were Greek, 4 were French, 2 were Gaelic, 1 was Aramaic, and 1 was Phoenician. No pronunciation guide was needed since all the names were familiar.

Phonotactic descriptions

The nineteenth-century name containing the most phonemes was *Alexander*, with ten. The names containing the least were *Roy* and *Earl*, with two each. *Milton*, *Paul*, and *Solomon* were the most like Standard American English, with average phoneme probabilities of .0802, .0729, and .0716. *Asa*, *Hugh*, and *Isaac* were the least, with average phoneme probabilities of .0128, .0180, and 0189. When bifones were considered, the names that were most like the English in the corpus were *Carl* and *Warren*, with average bifone probabilities of .0113 and .0108. The names that were the least like the English in the corpus were *Hugh* and *Earl*, both with an average bifone probability of .0000, and also *Roy* with .0001. Overall, longer names were more like English in the Brown Corpus.

Results of phonotactic comparisons

This study made a phonemic comparison of 55 male names from the fictional Spalding manuscript, 197 male names from *The Lord of the Rings*, and 100 male names from the nineteenth-century census records. A one-way ANOVA was used to compare the three groups of names on their average word length. The nineteenth-century census names (an average of 5.01 phonemes per name) were on average shorter than the Spalding names (5.91 phonemes), which were shorter than *Lord of the Rings* names (5.97 phonemes). Although the results were statistically significant ($F[2, 349] = 15.84$, $p < .0001$), the effect was not strong ($R^2 = .083$).

Statistically significant differences were also found among the three sources in average phoneme probability ($F[2, 349] = 5.57$, $p = .0042$) compared to the Brown Corpus, but again the effects were not strong ($R^2 = .031$). The average probability values for phonemes for names from the Spalding manuscript, *Lord of the Rings*, and nineteenth-century census names were .2887, .2591, and .2344 respectively. Differences in bifone probabilities across the three sources were not significant. A more detailed analysis comparing the three sources in their patterns of probabilities across the successive phoneme ordinal positions was needed.

An analysis of pattern was completed to examine how the three name sources compared at each ordinal position. In particular, we tested four hypotheses:

1. The overall variance of phonemic probabilities will distinguish between natural naming systems and fictional ones.
2. Fictional naming systems will have more consistency than natural naming systems in the mean phonotactic probabilities of names of varying word lengths.
3. Natural and fictional naming systems will differ in the multivariate patterns of mean phonotactic probabilities at each ordinal position.
4. Distributional properties will distinguish between natural and fictional systems, with natural being more Gaussian.

First hypothesis

The first hypothesis held that natural naming practices would show greater variance in all phonemic probabilities and also in all bifone probabilities than would fictional name systems, with the rationale that names created by single authors would be expected to be more similar in their phonotactic probabilities than names developing from a variety of origins within a natural language population. US Census names represented a natural name distribution. The Spalding manuscript was a fictional name system. Although *Lord of the Rings* was also fictional, the author's competence with ancient languages led us to expect more sophisticated name creation processes.

The natural naming system showed the greatest variance in phoneme probabilities, followed by *Lord of the Rings*, and finally the Spalding manuscript. The 100 names from the nineteenth century had a variance of phonemic probabilities (combining across the phoneme ordinal positions) of .01189. In contrast, the 197 names from *Lord of the Rings* had a variance of .01030, and the 55 names from the Spalding manuscript had a variance of .00299. The *F* ratio used to test for statistical significance for the comparison of nineteenth-century names to *Lord of the Rings* names was not significant. The comparison of nineteenth-century names to the Spalding manuscript names was significant ($F[99, 54] = 3.97$, $p = .00000009$), and the comparison of *Lord of the Rings* names to Spalding manuscript names was as well ($F[196, 54] = 3.45$, $p = .0000003$).

The pattern found in an analysis of the variances of bifone probabilities differed somewhat in that the nineteenth-century male names and names from *Lord of the Rings* were significantly different from one another ($F[99, 196] = 1.39$, $p = .0335$). The variances of probabilities for Spalding manuscript bifones (.00006) was about one-third as large as that for *Lord of the Rings* (.00018) and half as large as that for nineteenth century (.00013). The *F* ratio comparison of nineteenth-century names to the Spalding manuscript names was significant ($F[99, 54] = 2.06$, $p = .0021$), as was the comparison of *Lord of the Rings* names to Spalding manuscript names ($F[196, 54] = 2.87$, $p = .000008$).

Phonotactic probabilities of both phonemes and bifones (each combined across ordinal position) were highly successful in differentiating between the Spalding manuscript fictional names and the nineteenth-century male natural names; however, only the bifones differentiated between Tolkien's fictional names and the natural names. The relationship was not strong, and the variance for *Lord of the Rings* was slightly larger than for nineteenth-century names. We determined that more detailed analyses of these patterns of probabilities across ordinal positions were needed. We compared the three name sources at each ordinal position individually.

Since the Spalding manuscript had no names with fewer than four phonemes or more than eight, we restricted ourselves to this range and applied the telescoping method reported for the English vs. Latin pronunciation analysis of this manuscript to compare names of different lengths. Results are reported here for only the analysis of four or more phonemes, since the other analyses gave similar results. The analysis of four ordinal positions included 323 total male names: 55 from the Spalding manuscript, 179 from *The Lord of the Rings*, and 89 from the nineteenth-century census records.

Figure 2 shows the variances of phonotactic probabilities for each of the three name sources at each of four ordinal positions. The nineteenth-century names and *Lord of the Rings* names were both fairly similar, but the Spalding manuscript had significantly lower variances in phonotactic probabilities at ordinal positions one, two, and three, but not four. Because the variances were highest for the nineteenth-century names, they became the numerator for testing variance ratios for statistical significance at each ordinal position. The Spalding variances of phonotactic probabilities differed significantly from those of the nineteenth century in the first three ordinal positions ($F[88, 54] = 1.58$, $p = .0355$; $F[88, 54] = 2.20$, $p = .0011$; $F[88, 54] = 1.74$, $p = .0149$; $F[88, 54] = 0.94$, n.s.). However, those for *Lord of the Rings* did not differ appreciably from those of the nineteenth century ($F[88, 178] = 1.24$, n.s.; $F[88, 178] = 0.94$, n.s.; $F[88, 178] = 1.00$, n.s.; $F[88, 178] = 1.44$, $p = .0206$).

Hypothesis 1 was also tested for the variances of phonotactic probabilities of bifones. The results are shown in Figure 3. As with the phonemes, the nineteenth-century and *Lord of the Rings* names were very similar and the Spalding manuscript varied substantially. In the first bifone position, the Spalding manuscript names differed significantly from nineteenth-century names ($F[88, 54] = 2.18$, $p = .0013$), and from *Lord of the Rings* names ($F[178, 54] = 2.12$, $p = .0009$). In the second bifone position the names from *Lord of the Rings* and nineteenth century did not differ from each other, but both had a much higher variance than Spalding ($F[178, 54] = 2.87$, $p = .000009$; $F[88, 54] = 2.76$, $p = .00005$). On Bifone 3, none of the variances differed significantly.

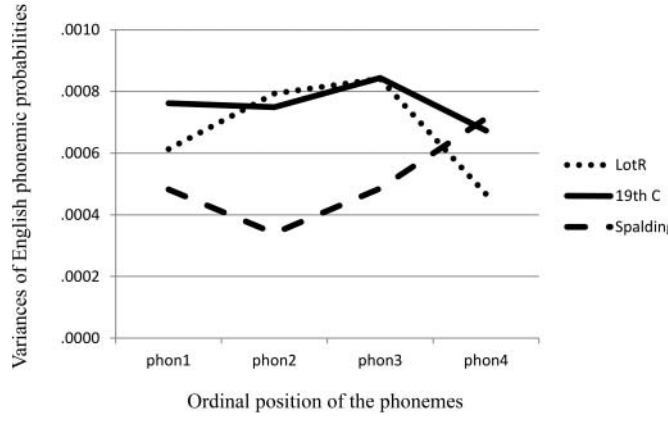

FIGURE 2 The variances of English phonemic probabilities for each of the three sources at each of the four ordinal phoneme positions.

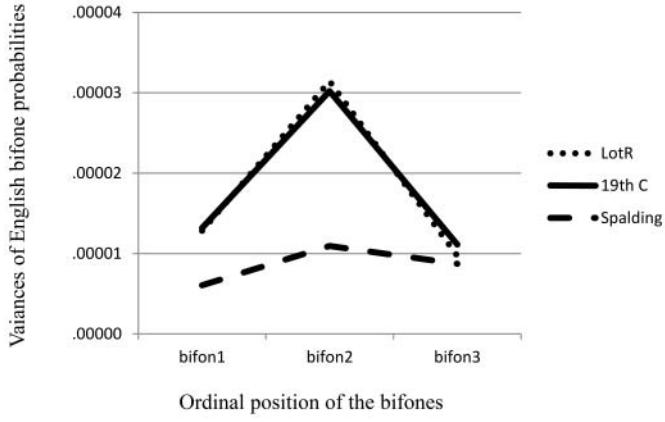

FIGURE 3 The variances of English bifone probabilities for each of the three sources at each of the four ordinal bifone positions.

Hypothesis 1 held that the variance of phonemic probabilities would distinguish between natural naming systems and fictional ones. While all of the tests differentiated clearly between the Spalding names and the natural nineteenth-century names, the fictional names crafted by Tolkien were similar to natural naming patterns.

Second hypothesis

The second hypothesis was extended from the first variance hypothesis, but was more subtle, dealing with variances of mean phonotactic probabilities for names at various word lengths. We assumed that an author's artificial naming system would show the same processes regardless of name length, but that names chosen by people from varying origins and backgrounds would involve a more heterogeneous set of phonotactic structures. The phonotactic probabilities of individual phonemes in natural naming practice would therefore be expected to vary more across name lengths.

This was tested with a two-way multivariate analysis of variance of the interactive effects of name source and name length on the ordinal position profiles of phonotactic probabilities. The results from four multivariate tests (Wilks' lambda, Pillai's trace, the Hotelling-Lawley statistic, and Roy's greatest root) are shown in Table 1 for each of the three sources of variance (name source, name length, and interaction between these two). All of these 12 multivariate tests were statistically significant, indicating that phonotactic ordinal position profiles were predictable from length, source, and the interaction of the two. Name source was a moderating variable, while name length showed differing effects for each source.

A similar two-way MANOVA was run on the bifone data. The multivariate tests for name source were not statistically significant. The effects of name length were significant (Wilks' lambda [4, 308] = 0.92693, p = .0248), as were the interaction effects (Wilks' lambda [8, 308] = 0.84942, p = .0012). The pattern of results from these parallel analyses on bifones gave similar results and will therefore not be reported here.

Figure 4 shows the simple effects of name length (from a length of four to a length of eight) on profiles of phonotactic probability over four ordinal positions. The

TABLE 1

THE FOUR MULTIVARIATE STATISTICS FOR A TWO-WAY MANOVA OF PHONEME PROBABILITIES AS A FUNCTION OF SOURCE OF THE NAMES AND NAME LENGTH

variance source	test	value	F	num df	den df	*p*
namesource	Wilks'	.9428	2.28	8	610	.0209
	Pillai's	.0580	2.28	8	612	.0206
	H-L	.0598	2.28	8	433.39	.0215
	Roy's	.0380	2.90	4	306	.0221
length	Wilks'	.8875	2.32	16	932.43	.0023
	Pillai's	.1151	2.28	16	1232	.0027
	H-L	.1238	2.35	16	604.03	.0021
	Roy's	.0947	7.29	4	308	<.0001
NxL interaction	Wilks'	.8368	1.74	32	1126.4	.0067
	Pillai's	.1708	1.72	32	1232	.008
	H-L	.1861	1.77	32	787.07	.0059
	Roy's	.1231	4.74	8	308	<.0001

profile at the top, for names of eight phonemes, had the overall highest English probabilities. The profile for names of length four appeared at the bottom, with profiles for name lengths of five, six, and seven generally lining up in order between these two. The figure also shows that the second phonemic position had higher English probabilities than the other three positions.

The univariate tests shown in Table 2 followed-up on the holistic multivariate results given in Table 1, breaking down the overall statistical significance of profile comparisons into which particular ordinal positions most accounted for significance. All three sets of multivariate tests showed statistical significance for the overall gestalt forms, the holistic patterns, and ten out of the twelve corresponding univariate tests

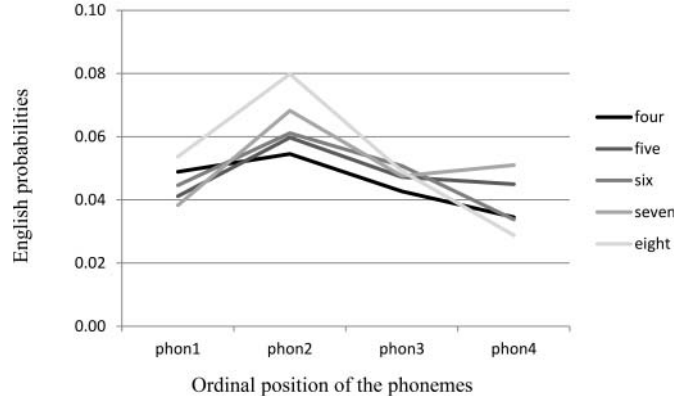

FIGURE 4 A comparison of the English phonemic probabilities for each word length at each ordinal position for all three name sources combined.

TABLE 2

FOUR TWO-WAY ANOVAS FOR A TWO-WAY MANOVA OF PHONEME PROBABILITIES AS A FUNCTION OF SOURCE OF THE NAMES AND NAME LENGTH (ONE ANOVA FOR EACH OF THE FOUR PHONEME POSITIONS)

ordinal position	variance source	df	SS (Type 3)	F	*p*	*R2*
Phoneme 1	namesource	(2, 308)	.0051	4.20	.0158	.0247
	length	(4, 308)	.0028	1.15	.3348	.0135
	NxL interaction	(8, 308)	.0135	2.80	.0053	.0657
	entire model	(14, 308)	.0196	2.32	.0048	.0952
Phoneme 2	namesource	(2, 308)	.0033	2.49	.0844	.0139
	length	(4, 308)	.0051	1.93	.1055	.0216
	NxL interaction	(8, 308)	.0068	1.28	.2523	.0287
	entire model	(14, 308)	.0331	3.56	<.0001	.1394
Phoneme 3	namesource	(2, 308)	.0012	0.80	.4524	.0050
	length	(4, 308)	.0012	0.38	.8246	.0047
	NxL interaction	(8, 308)	.0064	1.02	.4200	.0255
	entire model	(14, 308)	.0099	0.90	n.s.	.0395
Phoneme 4	namesource	(2, 308)	.0024	2.25	.1068	.0127
	length	(4, 308)	.0107	4.96	.0007	.0566
	NxL interaction	(8, 308)	.0092	2.14	.0317	.0483
	entire model	(14, 308)	.0252	3.34	<.0001	.1319

were also significant. Graphical summaries of the patterns of means were needed to understand the basis of the significant results.

Figures 5 and 6 show the phonotactic probability profiles as a function of name length for nineteenth-century names and for *Lord of the Rings* names. Both of these followed a similar pattern peaking in the English probabilities at phoneme two, but the nineteenth-century names had substantially more phonotactic variation than did the names from *Lord of the Rings*, particularly in the first ordinal position and in the fourth. The variance of the mean phonotactic probabilities in the first ordinal position for nineteenth-century names was 0.288 compared to only 0.018 for *Lord of the Rings* ($F[4, 4] = 15.77$, $p = .0044$). For the fourth ordinal position the variance of nineteenth-century names was 0.222 compared to only 0.011 for *Lord of the Rings* ($F[4, 4] = 21.13$, $p = .0022$).

Figure 7 shows the phonotactic probability profiles for the Spalding names. English probabilities peaked at phoneme two like the others, but the range of probabilities at phoneme positions one and four was more comparable to those of nineteenth-century names than *Lord of the Rings* names. The variance of the mean phonotactic probabilities in the first ordinal position for the Spalding manuscript was 0.255 compared to 0.288 for nineteenth-century names ($F[4, 4] = 1.13$, n.s.). For the fourth ordinal position the variance of Spalding manuscript names was 0.207 compared to 0.222 for nineteenth-century names ($F[4, 4] = 1.07$, n.s.). Although the Spalding manuscript names were comparable to the nineteenth-century names in variances of

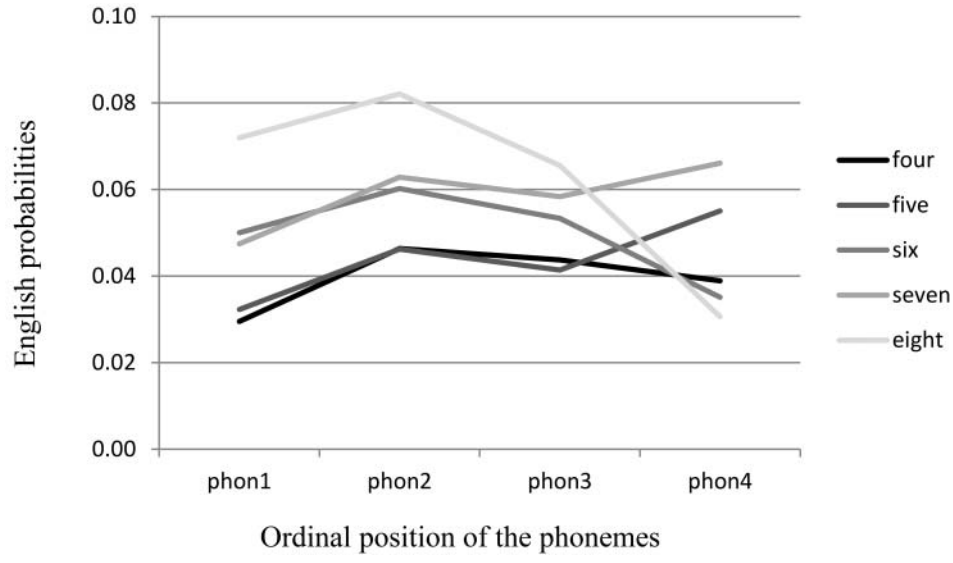

FIGURE 5 A comparison of the English phonemic probabilities for each word length at each ordinal position, names from the 19th century.

mean probabilities, the patterns of the probability profiles shown in Figure 7 were much less orderly than the patterns in Figures 5 and 6, which showed a clear trend for longer names to be more English like.

Figure 8 shows the variances of mean phonotactic probabilities for each of the three sources at each phonemic ordinal position. The pattern of variances for Spalding

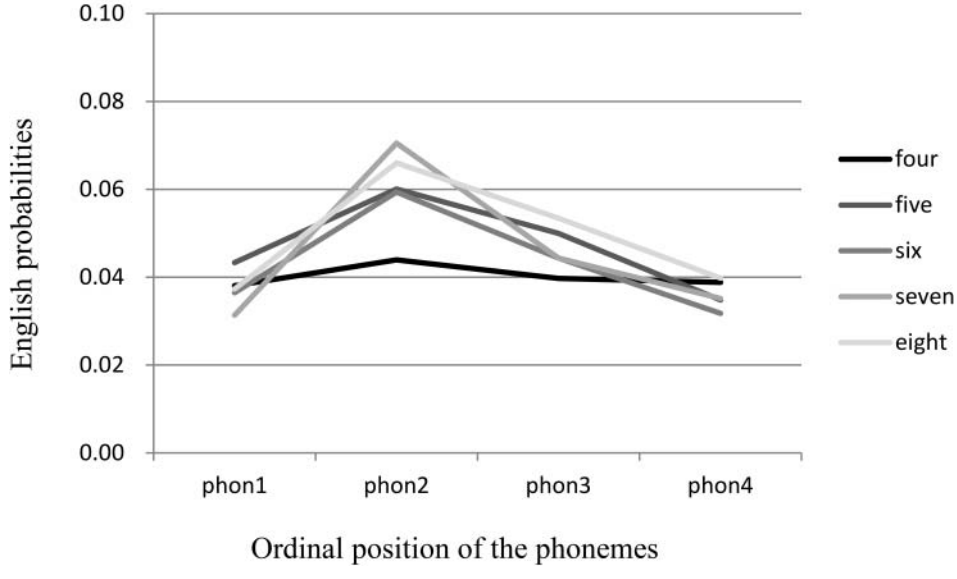

FIGURE 6 A comparison of the English phonemic probabilities for each word length at each ordinal position, names from *Lord of the Rings*.

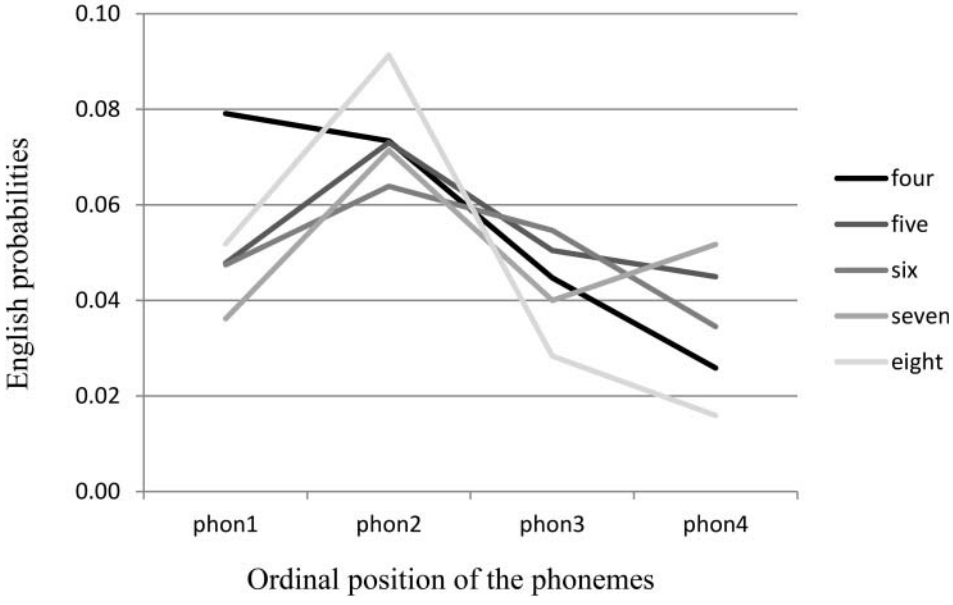

FIGURE 7 A comparison of the English phonemic probabilities for each word length at each ordinal position, names from the Spalding manuscript.

names and nineteenth-century names was U-shaped, whereas the pattern for *Lord of the Rings* names was inverted.

The tests of the first hypothesis showed a clear separation between the natural names and the Spalding names, but not *Lord of the Rings* names. The tests of the second hypothesis showed clear evidence of the separation of the natural names from the *Lord of the Rings* names but not from the Spalding names.

Third hypothesis

Whereas the first hypothesis explained the variances of phonotactic probabilities and the second hypothesis was concerned with the variances of means within each ordinal

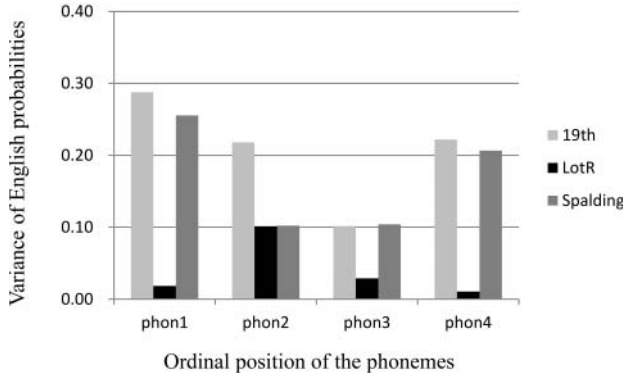

FIGURE 8 A comparison of the variances across the five word lengths of names for each of the three sources, at each of the first four ordinal positions.

position, the third hypothesis deals with the patterning of the mean probabilities themselves. It held that natural and fictional naming systems would differ in the multivariate patterns of mean phonotactic probabilities across the ordinal positions. We employed a one-way multivariate analysis of variance (MANOVA with contrasts) to test the pattern match of means across the three sources, with source as the independent variable and phonotactic probabilities at each ordinal position as the dependent variable. Although we conducted a similar one-way MANOVA analysis with bifone ordinal positions as the dependent variable, none of the statistical tests reached significance. Therefore, we report the results only for phonemes.

Tables 3 and 4 show the results of the one-way MANOVA and its associated ANOVAs for phonemes.[1] Both tables report three analyses: the significance level for the entire model comparing the three sources, the significance level for Contrast A of the two fictional naming systems compared to nineteenth-century names, and the significance level for Contrast B of *Lord of the Rings* compared to Spalding.

The results from four multivariate tests are shown in Table 1. The three sets of multivariate tests were each statistically significant.

In Contrast A of Table 1, natural naming systems (nineteenth century) were differentiated from fictional naming systems. The profile for the unsophisticated Spalding names differed significantly from that for the relatively sophisticated system in *Lord of the Rings*.

No significant model differences were found at the third phoneme position, but the entire model was significant in all other phoneme positions. Phoneme two was the only position in which all three tests were significant, as evident in the pattern shown in Figure 9, where phoneme two was the point where the three curves were most differentiated. The mean phonemic probability for Spalding (.070) was higher on phoneme two than that of *Lord of the Rings* (.060) which was higher than nineteenth-century names (.051).

TABLE 3

THE FOUR MULTIVARIATE STATISTICS FOR A ONE-WAY MANOVA OF PHONEME PROBABILITIES AS A FUNCTION OF NAME SOURCE (RESULTS FOR ENTIRE MODEL AND FOR TWO LINEAR CONTRASTS)

Variance Source	Test	Value	*F*	Num df	Den df	*p*	*Multivariate R2*
Entire Model	Wilks'	.8891	4.80	8	634	<.0001	.1109
	Pillai's	.1137	4.79	8	636	<.0001	
	H-L	.1216	4.81	8	450.54	<.0001	
	Roy's	.0843	6.70	4	318	<.0001	
Contrast A: Natural vs. Fictional	Wilks'	.9225	6.66	4	317	<.0001	.0775
	Pillai's	.0775	6.66	4	317	<.0001	
	H-L	.0840	6.66	4	317	<.0001	
	Roy's	.0840	6.66	4	317	<.0001	
Contrast B: LotR vs. Spalding	Wilks'	.9615	3.17	4	317	.0142	.0385
	Pillai's	.0385	3.17	4	317	.0142	
	H-L	.0400	3.17	4	317	.0142	
	Roy's	.0400	3.17	4	317	.0142	

TABLE 4

FOUR ONE-WAY ANOVAS SHOWING RESULTS FOR THE ENTIRE MODEL AND FOR THE TWO LINEAR CONTRASTS (ONE ANOVA FOR EACH OF THE FOUR PHONEME POSITIONS)

Ordinal Position	Variance Source	df	SS	*F*	*p*	*R2*
Phoneme 1	Contrast A: Natural vs. Fictional	(1, 320)	.0012	1.92	.1664	.0059
	Contrast B: LotR vs. Spalding	(1, 320)	.0036	5.67	.0178	.0174
	Entire Model	(2, 320)	.0039	3.07	.0480	.0188
Phoneme 2	Contrast A: Natural vs. Fictional	(1, 320)	.0111	15.68	<.0001	.0465
	Contrast B: LotR vs. Spalding	(1, 320)	.0038	5.46	.0201	.0162
	Entire Model	(2, 320)	.0120	8.51	0.0002	.0505
Phoneme 3	Contrast A: Natural vs. Fictional	(1, 320)	0.0001	0.07	.7918	0.0002
	Contrast B: LotR vs. Spalding	(1, 320)	0.0003	0.36	.5469	.0011
	Entire Model	(2, 320)	0.0003	0.19	.8312	.0012
Phoneme 4	Contrast A: Natural vs. Fictional	(1, 320)	.0056	9.99	.0017	.0296
	Contrast B: LotR vs. Spalding	(1, 320)	.0015	2.64	.1051	.0078
	Entire Model	(2, 320)	.0099	8.77	0.0002	.0520

Names from fictional sources were significantly different than natural names from the nineteenth century (Contrast A) in the second and fourth phoneme positions. Names from Spalding were significantly different from names from *Lord of the Rings* (Contrast B) in the first and second phoneme positions. Overall, Spalding names were most English-like and nineteenth-century names were least English-like.

Hypothesis 3 held that natural and fictional naming systems would be differentiated by the multivariate patterns of mean phonemic probabilities. Table 3 shows strong and significant effects for Contrast A, natural versus fictional, accounting for about 8% of the variance.

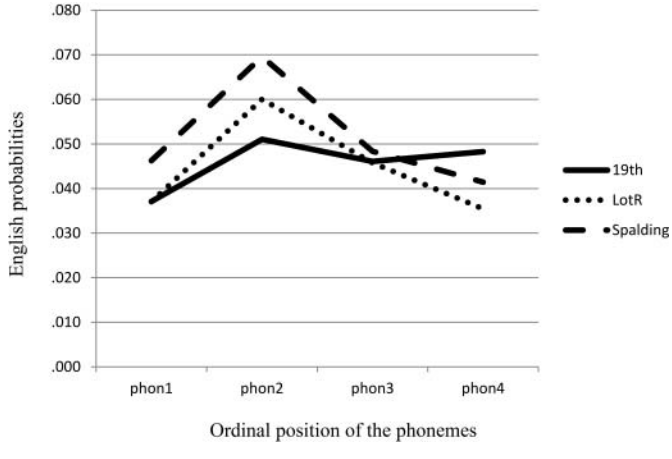

FIGURE 9 A comparison of the three name sources in their average English phonemic probabilities at each of the four ordinal positions.

Fourth hypothesis

The fourth hypothesis held that distributional properties could also distinguish between historical naming systems and individual author systems. Tests for normality (Gaussian shape of the data distributions) were of particular interest as descriptive measures for differentiating the three sources.

The twelve name distributions were tested for fit to a Gaussian curve using the Shapiro-Wilk test of normality (Shapiro and Wilk, 1965), and the Shapiro-Francia test of normality (Shapiro and Francia, 1972). They were also tested graphically for fit using the Q-Q plot. In the second phoneme position, the natural nineteenth-century names had a relatively Gaussian distribution, both in the numerical tests and in the graphical Q-Q plot — not surprising, as the second phoneme position is often a vowel in the center of the first morpheme within a name. The other three phoneme positions for nineteenth-century names had statistically significant departures from Gaussian shape. For *Lord of the Rings* all four phoneme positions departed significantly from normality, and for the Spalding manuscript all did but phoneme position three.

Figure 10 compares the Q-Q plots for each of the three sources in the second ordinal position. The Y axis of this plot shows the quantile values from the distribution in question, and the X axis shows the corresponding quantile values for a normal distribution. When the data are normally distributed, the plotted bivariate points form a straight diagonal line from the bottom left of the figure to the upper right. Superimposed upon the Q-Q plots for each name source are the results of the tests of normality.

The repetition of one particular phoneme in Spalding names, and two in *Lord of the Rings* are obvious. These were a large part of the basis for the departure in normality in the two distributions. For Spalding, the particularly common phoneme was /æ/ as in *cat* or *black*; 30 of the 55 names in the Spalding manuscript (54.6%) used this phoneme in the second position (e.g., Bambo, Baska, Fabious, Gamasko, Habelan, Hamack, Dato, Lambon, Rambock, Sambol). Tolkien's two most common phonemes in the second ordinal position were /r/ as in *red* and *try*, and /oʊ/ as in *go* and *home*; 42 of the 179 names in *Lord of the Rings* (23.5%) used /r/ in the second position (e.g., Aragorn, Brandir, Bregor, Frodo, Grishnak), and 21 of the 179 (11.7%) used /oʊ/ in the second position (e.g., Boromir, Lorien, Nori, Roac). Repetition patterns of this type were not so obvious in the more Gaussian pattern for the nineteenth-century names.

These differences between the natural naming pattern of the nineteenth century and the patterns for the two fictional systems are brought into perspective by comparing the percentage of names accounted for by the two most common phonemes in the second ordinal position and also in the other three positions as shown in Figure 11.

In both fictional name systems, the percentage of names accounted for by the two most common phonemes became less for each successive ordinal position from the second phoneme on, whereas for the nineteenth-century names the percentage increased with each successive ordinal position. There was a statistically significant difference in profile trends of the Figure 11 percentages between natural names and both of the fictional naming systems. For Spalding names compared to nineteenth-century names the chi square value was found to be large ($\chi^2(3) = 24.417$, $p < .0001$).

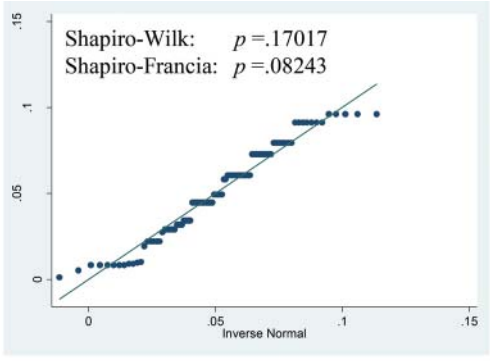

FIGURE 10 Q-Q plots for examining departure from Gaussian distributional shape: A comparison of the three manuscript sources at the second phoneme position. Also shown are the *p* values for the Shapiro-Wilk and the Shapiro-Francia tests of normality.

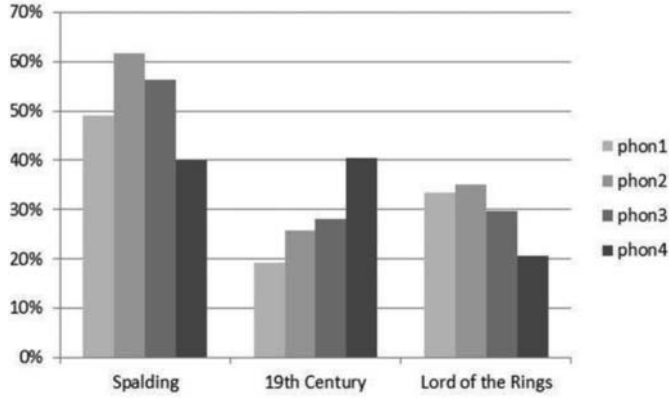

FIGURE 11 Bar graphs showing the percents of names from each of the three sources accounted for by the two most common phonemes in each of the four ordinal positions.

It was somewhat smaller for *Lord of the Rings* names compared to nineteenth-century names, but still significant ($\chi^2(3) = 11.360$, $p < .0099$). The profile trend of *Lord of the Rings* names did not differ significantly from that for Spalding names ($\chi^2(3) = 0.923$, $p < .8198$), indicating clear differentiation between natural and fictional naming patterns.

Conclusion

The four hypotheses were selected to test whether natural naming systems could be distinguished from fictional ones, first by overall variance of phonemic probabilities, second by the variance in mean probabilities across word lengths at each ordinal position, third by multivariate profiles of mean phonotactic probabilities, and fourth by the distributional properties of the phonemic probabilities. The test of the first hypothesis distinguished Spalding names from natural names, but *Lord of the Rings* names were not distinguished from natural. In contrast, the test of the second hypothesis distinguished *Lord of the Rings* names from natural, but not Spalding names. However, on both Hypothesis 3 and Hypothesis 4, both of the fictional sources of names were clearly distinguished from natural names.

This research was an exploratory study to determine whether there were sufficient phonemic differences between fictional and authentic names to merit further investigation. Results indicated that there may be a phonoprint of sorts in the fictional work of both Tolkien and Spalding that may also surface in the work of other authors. Many more fictional and natural names from a variety of authors and time periods will need to be analyzed before final conclusions can be drawn.

Results of this study indicated that while it is possible to create a convincing set of names for a story, as Spalding and Tolkien did, such names seemed to follow patterns at the phoneme and bifone levels that were significantly different than those of authentic lists of names from a variety of cultural origins. The possibility of phonoprints invites further investigation, and the methods of analysis used in this study may lead to new ways of doing so.

Note

[1] Both ANOVA and MANOVA assume a population that is normally distributed (Gaussian). However, like the *t* test from which it is derived, ANOVA is robust with respect to violation of this assumption (Box, 1953), particularly when sample size is large, as it was here. Note in the discussion of Hypothesis 4 in the next section, however, that Gaussian shape of phoneme probabilities might be useful as a differentiator between natural naming systems and fictional ones.

Bibliography

Archer, John B., John L. Hilton, and G. Bruce Schaalje. 1997. "Comparative Power of Three Author-Attribution Techniques for Differentiating Authors." *Journal of Book of Mormon Studies* 6(1): 47–63.

Bailey, Todd M. and Ulrike Hahn. 2001. "Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods?" *Journal of Memory and Language* 44(4): 568–591.

Baily, Richard W. 1979. "Authorship Attribution in a Forensic Setting." *Advances in Computer-Aided Literary and Linguistic Research*: 9.

Box, George Edward Pelham. 1953. "Non-Normality and Test on Variance." *Biometrika* 40: 318–335.

Chance, Jane. 2004. *Tolkien and the Invention of Myth: A Reader*. Frankfort, KY: University of Press of Kentucky.

Croft, D. James. 1981. "Book of Mormon 'Wordprints' Reexamined." *Sunstone* 6(2): 15–22.

Dell, Gary S., Kristopher D. Reed, David R. Adams, and Antje S. Meyer. 2000. "Speech Errors, Phonotactic Constraints, and Implicit Learning: A Study of Experience in Language Production." *Journal of Experimental Psychology: Learning, Memory, & Cognition* 26(6): 1355–1367.

Downey, Sean S., Brian Hallmark, Murray P. Cox, Peter Norquest, and J. Stephen Lansing. 2008. "Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction." *Journal of Quantitative Linguistics* 15(4): 340–369.

Eldredge, J. Lloyd. 2005. *Teach Decoding: Why and How* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Erickson, Whitney, Angel L. Venegas, Alexandra Brattos, Malvania Salash, Donald W. Walker, Jessica H. Scott, and Bruce L. Brown. 2008. "Convergent Multivariate Graphics for Capturing the Naming Patterns in US Census Data from the Nineteenth Century." *Proceedings of the Section on Government Statistics*. Alexandria, VA: American Statistical Association.

Francis, W. Nelson. 1981. "Word-Making: Some Sources of New Words." *Language: Introductory Readings*. Ed. Virginia P. Clark, Paul A. Sschholz, and Alfred F. Rosa. New York: St Martin's Press, 316–328.

Hanks, Patrick and Flavia Hodges. 1995. *A Dictionary of First Names*. New York: NY: Oxford University Press.

Hilton, John L. 1990. "On Verifying Wordprint Studies: Book of Mormon Authorship." *BYU Studies* 30(3): 89–108.

Holmes, David I. 1994. "Authorship Attribution." *Computers and the Humanities* 28(2): 87–106.

Iqbal, Farkhund, Liaquat A. Khan, Benjamin C. M. Fung, and Mourad Debbabu. 2010. "E-mail Verification of Forensic Investigation." Available at: <http://users.encs.concordia.ca/~fung/pub/IKFD10sac.ppt> (Accessed 23 March, 2013).

Jusczyk, Peter W., Angela D. Friederici, Jeanine M. I. Wessels, Vigdis Y. Svenkerud, and Ann Marie Jusczyk. 1993 "Infants' Sensitivity to the Sound Patterns of Native Language Words." *Journal of Memory & Language* 33: 402–420.

Jusczyk, Peter W., Paul A. Luce, and Jan Charles-Luce. 1994. "Infants' Sensitivity to Phonotactic Patterns in the Native Language." *Journal of Memory & Language* 33(5): 630–645.

Kessler, Brett and Rebeca Treiman. 1997. "Syllable Structure and the Distribution of Phonemes in English Syllables." *Journal of Memory & Language* 37(3): 295–311.

Kucera, Henry and Nelson W. Francis. 1967. *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Morton, Andrew Q. 1979. *Literary Detection*. New York, NY: Scribner's Sons.

Rencher. A. C. 2002. *Methods of Multivariate Analysis* (2nd ed.). New York: John Wiley and Sons.

Shapiro, Samuel and R. S. Francia. 1972. "An Approximate Analysis of Variance Test for Normality." *Journal of the American Statistical Association* 67: 215–216.

Shapiro, Samuel and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52(3/4): 591–611.

Shih, Stephanie. 2012. "Linguistic Determinants of English Personal Name Choice." Paper presented at the annual conference of the Linguistic Society of America, Portland Oregon.

Starr, Rebecca. 2012. "Disambiguating Romanized Chinese Personal Names: A Corpus Based Approach to Backtransliteration and Gender Identification." Paper presented at the annual conference of the American Name Society, Portland, OR.

Storkel, Holly L. 2001. "Learning New Words: Phonotactic Probability in Language Development." *Journal of Speech, Language and Hearing Research* 44: 1321–1337.

Storkel, Holly L. 2003. "Learning New Words II: Phonotactic Probability in Verb Learning." *Journal of Speech, Language, & Hearing Research* 46: 1312–1323.

Vitevich, Michael S. 2002. "Influence of Onset Density on Spoken-word Recognition." *Journal of Experimental Psychology: Human Perception and Performance* 28: 270–278.

Vitevich, Michael S., David B. Pisoni, Karen I. Kirk, Marcia Hay-McCutcheon, and Stacey L. Yount. 2002. "Effects of Phonotactic Probabilities on the Processing of Spoken Words by Postlingually Deafened Adults with Cochlear Implants." *Volta Review* 102: 283–302.

Vitevich, Michael S. and Paul A. Luce. 2004. "A Web-based Interface to Calculate Phonotactic Probability for Words and Nonwords in English." *Behavior Research Methods, Instruments, & Computers* 36(3): 48–487.

Vitevich, Michael S. and Paul A. Luce. 1999. "Probabilistic Phonotactics and Spoken-word Recognition." *Journal of Memory & Language* 40(3): 374–408.

Vitevich, Michael S., Paul A. Luce, Jan Charles-Luce, and David Kemmerer. 1997. "Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words." *Language and Speech* 40: 47–62.

Whissell, Cynthia. 2001. "Sound and Emotion in Given Names." *Names* 49(2): 97–120.

Wright, Saundra K. 2012. "Naming Decisions Made by International Students Studying in the US". Paper presented at the annual conference of the American Name Society, Portland, OR.

Wu, Rui-Wen. 2010. "Development and Strata Analysis of Geng She Unrounded Cognates in Proto-Min." *Language and Linguistics* 11(2): 297–334.

Young, Steven. 2004. "'Old Prussian' in M. Pratorius's *Deliciae Prussicae*." *Studies in Baltic and Indo-European Linguistics: In Honor of William R. Schmalsteig*. Ed. Philip Baldi and Pietro U. Dini. Amsterdam, Netherlands: John Benjamin, 275–283.

Zheng, Rong, Yi Quin, Zan Huang and Hsinchun Chen. 2003. "Authorship Analysis in Cybercrime Investigation." Available at: <http://www.mendeley.com/catalog/authorship-analysis-cybercrime-investigation/> (Accessed 24 March, 2013).

Notes on contributors

Brad Wilcox is an Associate Professor in the Department of Teacher Education at Brigham Young University where he teaches graduate and undergraduate courses in literacy and children's literature. His research interests include reading, writing, education in international settings, and onomastics.

Correspondence to: Brad Wilcox, 201-P MCKB, Brigham Young University, Provo, UT 84602, USA. Email: brad_wilcox@byu.edu

Bruce L. Brown is a Professor in the Department of Psychology at Brigham Young University where he teaches graduate and undergraduate courses. His research interests include psychology, statistics, and the study of names.

Wendy Baker-Smemoe is an Associate Professor in the Department of Linguistics and English Language at Brigham Young University. Her research interests include second language acquisition, dialectology, and onomastics.

Sharon Black is an Associate Teaching Professor and editor/writing consultant in the David O. McKay School of Education at Brigham Young University. Her research interests include early literacy instruction, gifted/talented education, and elementary arts integration.

Justin Bray graduated from Brigham Young University in 2011 with a BA in history and a minor in Latin American studies. He currently works at the LDS Church History Library in Salt Lake City, Utah.