Frequency Analysis in Light of Language Innovation:

Exploring letter frequencies across time, from the days of Old English to the days of now.

Marsha Lynn Moreno
Math 187, Spring 2005

I. Introduction

Frequency tables and acronyms like "EAT ON IRS" are a common encounter for the novice cryptography student. They are a useful tool for analyzing messages encrypted by Caesar Cipher, Monoalphabetic Substitution, along with many other early cryptosystems. The facts are stated and the basics stick: what we find most often is assumed to be [e], and what we never find is most likely [x] or [z]. Such is the case for Modern English, but do these generalizations translate into Old English or Middle English? With centuries of years gone by, the English of today appears nothing like that of the 11th century or even the 14th century. Perhaps, a quick glance may persuade the reader. What follows is "The Lord's Prayer" (as stated in the Gospel of Luke) in Old English and Middle English:

> Cweðað þus. þonne ge eow gebiddað; Ure fæder þu ðe on heofone eart: si þin nama gehalgod tocume þin rice. gewurðe ðin willa on heofone and on eorþan. Syle us todæg urne dæghwamlican hlaf: And forgyf us ure gyltas. swa we forgyfað ælcum þara þe wið us agyltað: and ne læd þu us on costunge: ac alys us fram yfele. (Luke 11:2-4, the Polyglot Bible: English - 1000s, Mark Davies)

> Whan ÿee preÿen, sey ÿee, Fader oure, halewid be þi name, þi kingdam come to say ÿif to vs todai oure eche daÿis bred. Forÿif to vs oure synnes, as wee forÿiuen to eche owende to vs, leed vs not in to temptacioun. (Luke 11:2-4, the Polyglot Bible: English - 1300s, Mark Davies)

These texts are almost foreign to the eyes of Modern English speakers. Yet to the linguistically inclined, this is no surprise -- Old English is considered to be completely unintelligible from the language we speak today. This paper will touch on the linguistics of language, particularly how English has changed over time, and in what ways that innovation has affected frequency analysis for single letters in the language.

II. Methodology

While it had been my original intent to find three prominent authors for both the days of Old English and the days of Middle English, I found select passages of the Bible to be most available and accessible on the web. Indeed, the Bible speaks a language of its own, Christianese,

if you will. And this is bound to have some effect on our usual letter-frequencies for English. In fact, we shall soon see how even the Modern English translations of the Bible deviate from our standard frequency of letter usage, "EAT ON IRS."

This may seem problematic at first; however, if we are using the Bible to examine each version of English -- Modern, Middle, and Old -- then we will still be able to detect the changes in English over time, specifically for Biblical texts. By using the Bible for all purposes of this study, we will have less variability than we would otherwise. To create some balance between the Old Testament and the New Testament, I will use four Psalms from the Old Testament and four chapters of Luke for letter frequency counts. Now, you may have noticed already, from "The Lord's Prayer", that there exist letters in Old English and Middle English that no longer manifest in English today. I still account for these letters, [þ, æ, ð, ÿ], thereby adding four more letters to our twenty-six letter alphabet. So for Modern English frequencies, you will see zeros across the board for these letters. For Middle English, [þ, ÿ] will show up. For Old English, [þ, æ, ð] will *all* be present, while on the flip-side, [j, k] are ever absent from the language -- that is because these characters were not invented until much later.

III. Modern English Letter Frequencies

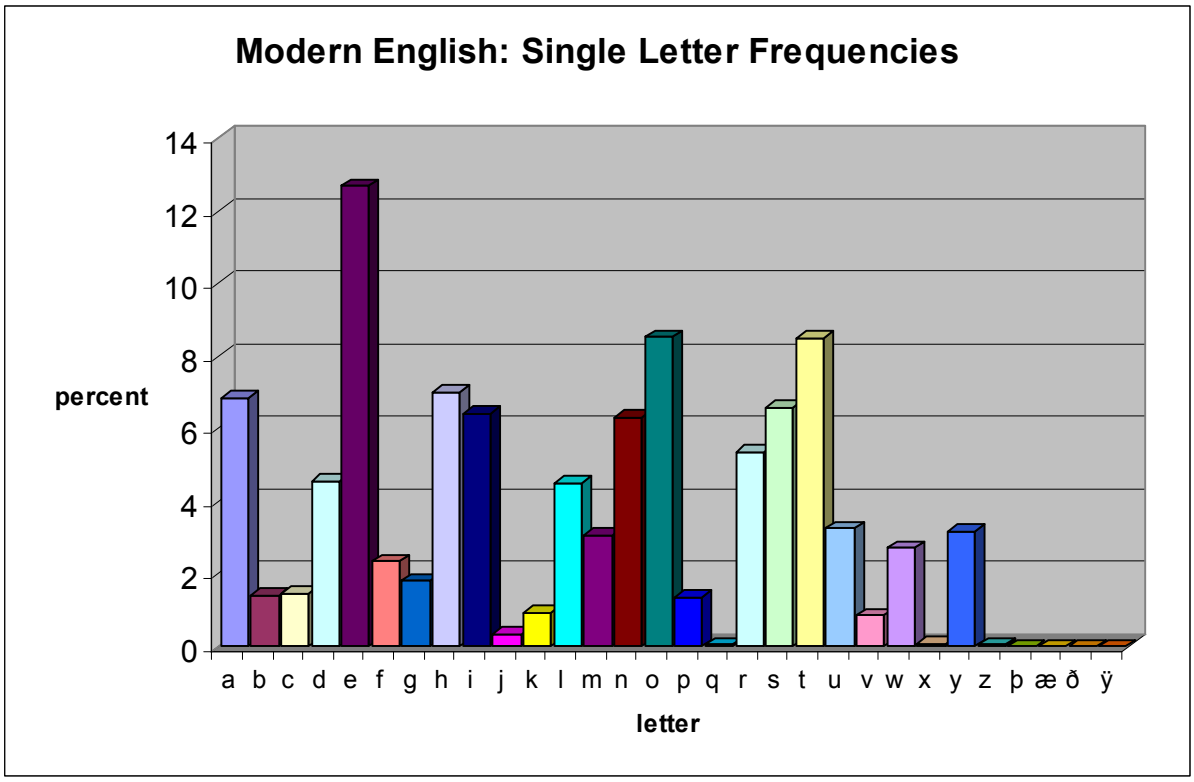
Since Modern English strikes familiar with us all, let us begin our frequency analysis here, noting the exact letter frequencies for the Bible passages I have selected and using their averaged values as reference points when we examine the Old English and the Middle English data.

ModE	Luke 5	Luke 9	Luke 14	Luke 24	Psalm 63	Psalm 107	Psalm 139	Psalm 143
total	3580	5393	3071	4010	756	2914	1575	953
a	7.91%	7.84%	7.49%	7.11%	5.95%	5.97%	7.56%	4.93%
b	1.37%	1.45%	1.95%	1.55%	2.12%	1.00%	1.14%	0.52%
c	1.87%	1.74%	1.63%	1.30%	1.06%	1.58%	1.33%	1.15%
d	4.44%	5.08%	3.74%	5.36%	2.91%	5.70%	5.02%	4.20%
e	13.12%	13.39%	12.86%	14.26%	10.71%	13.25%	12.25%	11.54%
f	2.31%	2.13%	2.15%	1.65%	2.25%	3.02%	2.48%	2.73%
g	2.03%	1.98%	1.56%	1.97%	1.59%	1.78%	1.71%	1.78%
h	7.21%	7.23%	6.81%	8.78%	6.75%	8.20%	6.41%	4.72%
i	6.03%	5.90%	7.10%	6.28%	6.61%	6.07%	5.78%	7.45%
j	0.61%	0.65%	0.29%	0.40%	0.53%	0.07%	0.00%	0.10%
k	0.92%	0.82%	0.59%	0.72%	0.93%	0.82%	1.40%	1.26%
l	3.85%	4.17%	4.53%	3.42%	7.41%	4.02%	4.06%	4.51%
m	2.68%	2.99%	2.47%	2.94%	2.65%	1.99%	4.13%	4.51%
n	7.37%	5.99%	6.94%	6.48%	4.37%	6.66%	6.86%	5.77%
o	7.93%	8.27%	8.50%	6.83%	9.39%	8.27%	8.57%	10.39%
p	1.42%	1.65%	1.47%	1.55%	1.72%	0.89%	0.95%	1.26%
q	0.03%	0.00%	0.16%	0.00%	0.00%	0.10%	0.06%	0.00%
r	4.50%	4.23%	4.23%	4.19%	5.56%	6.55%	5.97%	7.45%
s	6.76%	6.53%	6.64%	6.16%	7.67%	7.10%	4.70%	7.14%
t	9.19%	9.03%	9.57%	10.30%	6.35%	9.75%	7.43%	6.40%
u	2.40%	2.76%	3.03%	2.37%	4.37%	2.37%	3.94%	4.72%
v	0.56%	1.17%	1.04%	0.87%	0.66%	0.72%	0.63%	1.05%
w	3.35%	2.67%	2.54%	2.89%	2.91%	2.37%	3.05%	1.89%
x	0.11%	0.07%	0.26%	0.02%	0.00%	0.03%	0.06%	0.00%
y	1.90%	2.23%	2.44%	2.44%	5.56%	1.68%	4.44%	4.51%
z	0.08%	0.04%	0.00%	0.15%	0.00%	0.03%	0.00%	0.00%
þ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
æ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ð	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ÿ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

ModE	Luke avg	Psalm avg	Overall avg
total	4013.5	1549.5	2781.5
a	7.59%	6.10%	6.84%
b	1.58%	1.20%	1.39%
c	1.64%	1.28%	1.46%
d	4.66%	4.46%	4.56%
e	13.41%	11.94%	12.67%
f	2.06%	2.62%	2.34%
g	1.89%	1.72%	1.80%
h	7.51%	6.52%	7.01%
i	6.33%	6.48%	6.40%
j	0.49%	0.18%	0.33%
k	0.76%	1.10%	0.93%
l	3.99%	5.00%	4.50%
m	2.77%	3.32%	3.05%
n	6.70%	5.92%	6.31%
o	7.88%	9.16%	8.52%
p	1.52%	1.21%	1.36%
q	0.05%	0.04%	0.04%
r	4.29%	6.38%	5.34%
s	6.52%	6.65%	6.59%
t	9.52%	7.48%	8.50%
u	2.64%	3.85%	3.25%
v	0.91%	0.77%	0.84%
w	2.86%	2.56%	2.71%
x	0.12%	0.02%	0.07%
y	2.25%	4.05%	3.15%
z	0.07%	0.01%	0.04%
þ	0.00%	0.00%	0.00%
æ	0.00%	0.00%	0.00%
ð	0.00%	0.00%	0.00%
ÿ	0.00%	0.00%	0.00%

Most Frequent to Least Frequent: e o t h a s i n r d l u y m w f g c b p k v j q x z {þ, æ, ð, ÿ}

Note: "{ }" indicates that letters therein do not exist in the spelling system. In other words, the frequency for these letters is 0.00%.



Rather than "EAT ON IRS," we have "EOT HAS IN." So [e] is the most frequent, as expected. But [h] is showing up much more in the Bible than in President Bush's 2004 State of the Union speech, which (if you didn't know already) is where we get "EAT ON IRS." Besides [h] replacing [r] in the top eight, we have the same letters; they just happen to be in a slightly different order. Naturally, [e] is first place, while [t] maintains its rank as third.

IV. Time Warp to the 11th Century: Old English

i. Linguistic Structure and Predictions

Exhibiting more than just different letters in the alphabet, Old English diverges from Modern English in a number of linguistic areas: syntax (sentence structure), morphology (word structure and inflections), and phonology (sound patterns). I will touch briefly on some linguistic

structure so that the reader may have a better understanding as to why certain letters occur more or less in Old English than in Modern English.

First of all, while Modern English is an SVO (Subject-Verb-Object) language, Old English had two possible sentence constructions: SVO or SOV. The syntactic system for Modern English relies on very strict word order, but Old English had much more freedom in how words are arranged. Such is the case for Old English because it had a more fully developed system for morphological inflection on words.

We will look at case marking to help illustrate the more flexible word ordering in Old English. In Modern English, case marking scarcely exists, but one place in which we still find it is in our pronoun system. For first person pronouns, we have the following:

	Singular	Plural
Subjective:	I	we
Objective:	me	us
Possessive:	my (mine)	our (ours)

Whether one hears "I" or "me" or "we," he/she is able to figure out exactly what role that pronoun plays in a particular sentence. Given "I," it is predictable that the speaker refers to himself, and only himself, as the subject of the sentence. Given "us," we know that the speaker refers to himself and at least one other person and that some event takes place in their direction (e.g. "John hit us" and "John gave the book to us").

This instance of case marking for Modern English pronouns cannot even measure up to the case marking we find for the majority of nouns in Old English. Old English nouns were based on an inflectional system that divided them into four categories: nominative (what we know as subjective in Modern English), genitive (possessive), accusative (corresponds to direct objects), and dative (often corresponds to indirect objects). Since this type of information for the noun was

carried in the form of a suffix, there is less need to have specific word order. Below are the different case markings for the Old English noun, "king":

	Singular	Plural
Nominative:	cyning	cyningas
Genitive:	cyninges	cyningas
Dative:	cyninge	cyningum
Accusative:	cyning	cyningas

This type of inflection took place for, not all, but many of the nouns in Old English. So on the topic of cryptanalysis, we can see that [a] is being used as a grammatical marker in plural forms, whereas in Modern English, we are used to finding [e] in plural forms, as in the suffix -es. This part of the Old English grammar is likely to cause [a]'s frequency to increase. Another important aspect of case marking is that it rids the need for prepositions when referring to direct and indirect objects. Therefore, the preposition "to," which we often see in Modern English, will not be occurring in the contexts of direct objects and indirect objects for Old English. With that given, frequencies for [t] and [o] are likely to appear less in Old English than in Modern English.

In Old English, plural forms did not always call for the presence of some suffix but rather a change in the vowel within the word. We see this type of behavior in words like "tooth/ teeth" and "mouse/ mice." This vowel change within word roots is another common feature for both nouns and verbs in Old English. Therefore, a vowel change that indicates number will have some effect on the frequency of [s] since the general rule of today's English is "add -es (or just -s)" to make something plural.

Vowel changes often occurred when forming past tense, as seen in our modern-day *irregular verbs*: sing/ sang, wake/ woke, sit/ sat, hold/ held, and so on. But much of these verbs have been *regularized.* As their usage has become less frequent, their irregular past tense form

has been lost and replaced with the typical [-ed] suffix. For example, we are starting to hear both "abided" *and* "abode" for the past forms of "abide" because the usage of this verb is so low.

Besides the type of inflection where the vowel quality changes in the root, there were certainly verbs in Old English that were suffixed in order to indicate past tense. Observe the following inflections for word root "lufian" (*to love*):

	Singular Present	Plural Present	Singular Past	Plural Past
1st person:	ic lufie	we lufiað	ic lufode	we lufodon
2nd person:	þu lufast	ge lufiað	þu lufodest	ge lufodon
3rd person:	he lufað	hi lufiað	he lufode	hi lufodon

The [-don] suffix present in the Plural Past suggests a higher frequency for [n]. We should also note that the infinitive form of the verb is "lufian," where "to" is represented by the suffix [-ian]. Imagine all the instances in Modern English where we say "I want to *v*" or "I tried to *v,*" where *v* stands for some verb. For every "to *v*" in Modern English, there is a corresponding Old English verb containing either [-ian] or [-an] as an infinitive marker. For these reasons, it is safe to assume a much higher frequency, not only for the letter [n], but also for the letter [a].

Now, I will turn my focus to spelling conventions in Old English. I will particularly discuss the letters that no longer exist in Modern English, as well as their corresponding sounds. I have listed below the three characters used in Old English that are no longer seen in Modern English, in fact, the latter two had already dropped off by the time of Middle English:

[þ] -- called "eth;" pronounced as the "th" in "theater" or in "then"

[ð] -- called "thorn;" pronounced the same as "eth"

[æ] -- called "ash;" pronounced as the "a" in "cat"

Both [þ] and [ð] stood alone in Old English, representing the *two-letter* sound we have in Modern English, [th]. This will have a direct effect on the frequencies for [t] and [h]; these

particular characters will be least frequent in Old English, more frequent in Middle English (since [ð] drops out of the spelling system), and most frequent in Modern English (since [þ] drops out).

Based on properties of syntax, morphology, and spelling conventions, it is quite evident that the linguistic structure for Old English is far from that of Modern English. With these grammatical differences, we should see some deviance from the Modern English letter frequencies. Interestingly, we may find that some letters have maintained similar frequencies across time, despite all the language innovation English has undergone through the centuries.

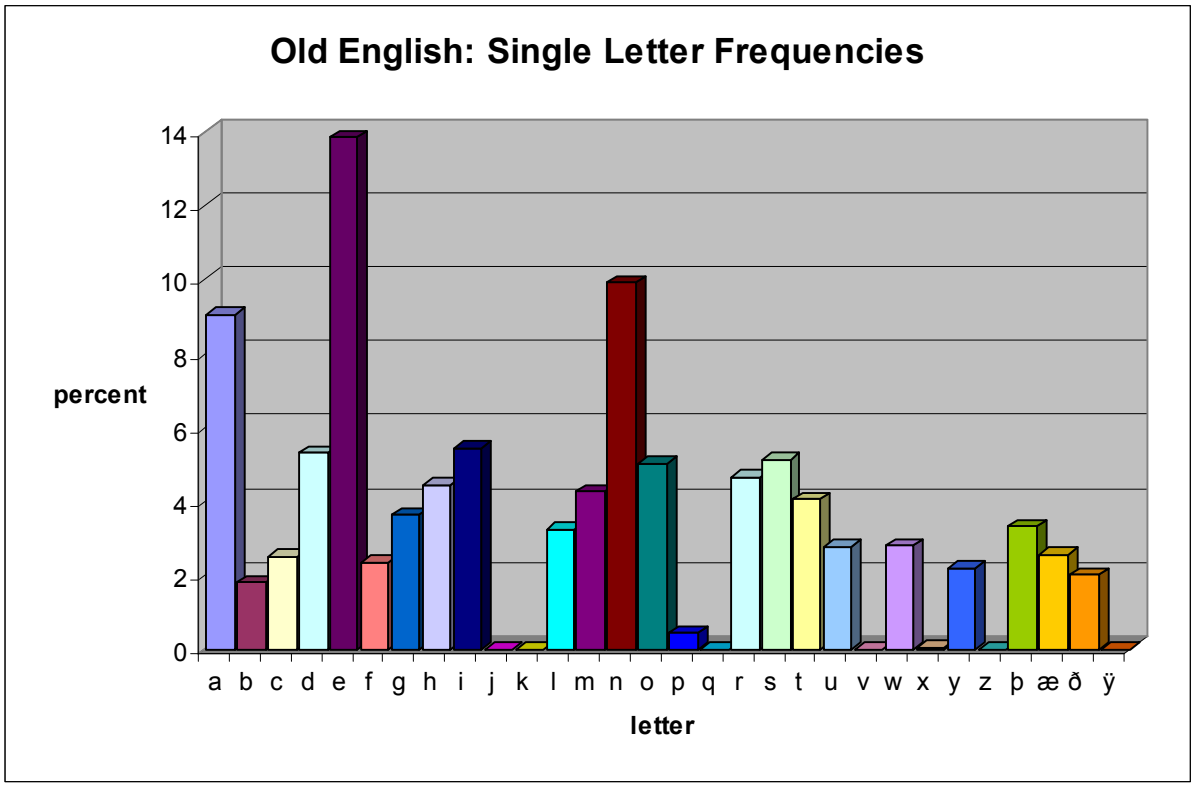
ii. Letter Frequencies

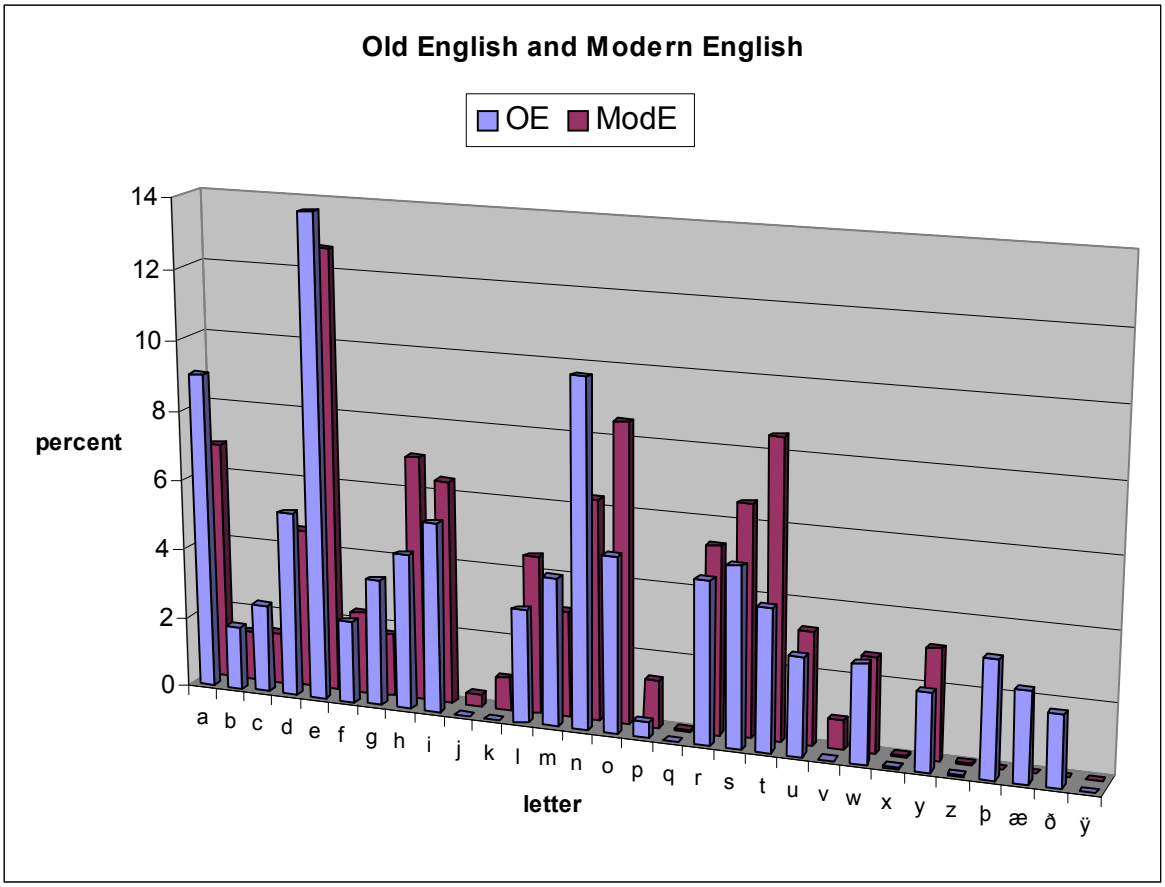
OE	Luke 5	Luke 9	Luke 14	Luke 24	Psalm 63	Psalm 107	Psalm 139	Psalm 143
total	3401	5092	2838	3822	800	1163	1302	1685
a	9.76%	9.23%	8.70%	9.60%	10.13%	7.82%	8.60%	8.49%
b	1.50%	1.22%	1.83%	1.41%	1.00%	0.69%	5.38%	1.31%
c	2.73%	2.20%	2.54%	1.96%	3.13%	3.35%	1.84%	2.43%
d	5.91%	5.75%	4.69%	7.04%	4.75%	4.82%	4.53%	4.93%
e	12.80%	13.45%	14.34%	12.95%	12.38%	15.65%	15.21%	13.77%
f	2.03%	2.42%	2.08%	1.86%	1.63%	2.67%	3.61%	2.37%
g	3.09%	3.55%	3.56%	4.58%	3.75%	3.53%	3.61%	3.26%
h	4.38%	5.44%	4.62%	4.97%	4.75%	3.44%	4.15%	3.80%
i	4.97%	5.18%	5.36%	5.29%	6.00%	5.33%	4.99%	6.29%
j	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
k	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
l	3.26%	2.95%	3.35%	2.64%	2.88%	3.96%	3.15%	3.62%
m	3.59%	3.44%	4.02%	3.38%	4.75%	4.64%	4.92%	5.28%
n	11.35%	10.13%	10.01%	11.02%	10.13%	9.46%	9.29%	7.77%
o	5.38%	5.50%	4.26%	5.42%	5.63%	5.33%	4.45%	4.21%
p	0.62%	0.53%	0.42%	0.42%	0.38%	0.26%	0.23%	0.59%
q	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
r	3.91%	3.79%	3.70%	4.19%	5.00%	5.16%	6.30%	5.04%
s	4.94%	5.75%	4.44%	4.47%	5.88%	6.10%	3.76%	5.64%
t	3.65%	3.55%	4.26%	3.61%	4.38%	3.35%	4.22%	5.28%
u	2.41%	2.67%	2.64%	2.43%	2.13%	3.18%	3.15%	3.56%
v	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
w	3.09%	3.10%	2.47%	3.09%	3.00%	1.98%	3.15%	2.43%
x	0.06%	0.08%	0.11%	0.02%	0.00%	0.00%	0.00%	0.06%
y	2.03%	1.85%	2.85%	2.04%	2.13%	2.58%	2.23%	1.84%
z	0.03%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%
þ	4.18%	3.91%	4.58%	3.32%	1.50%	2.75%	3.30%	2.97%
æ	3.29%	2.95%	2.68%	3.27%	1.63%	2.24%	2.15%	2.37%
ð	1.06%	1.36%	2.50%	0.99%	3.13%	1.72%	2.61%	2.67%
ÿ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

OE	Luke avg	Psalm avg	Overall avg
total	3788.25	1237.5	2512.875
a	9.32%	8.76%	9.04%
b	1.49%	2.10%	1.79%
c	2.36%	2.69%	2.52%
d	5.85%	4.76%	5.30%
e	13.39%	14.25%	13.82%
f	2.10%	2.57%	2.33%
g	3.70%	3.54%	3.62%
h	4.85%	4.04%	4.44%
i	5.20%	5.65%	5.43%
j	0.00%	0.00%	0.00%
k	0.00%	0.00%	0.00%
l	3.05%	3.40%	3.23%
m	3.61%	4.90%	4.25%
n	10.63%	9.16%	9.90%
o	5.14%	4.91%	5.02%
p	0.50%	0.37%	0.43%
q	0.00%	0.00%	0.00%
r	3.90%	5.38%	4.64%
s	4.90%	5.35%	5.12%
t	3.77%	4.31%	4.04%
u	2.54%	3.01%	2.77%
v	0.00%	0.00%	0.00%
w	2.94%	2.64%	2.79%
x	0.07%	0.02%	0.04%
y	2.19%	2.20%	2.19%
z	0.01%	0.00%	0.01%
þ	4.00%	2.63%	3.31%
æ	3.05%	2.10%	2.57%
ð	1.48%	2.53%	2.01%
ÿ	0.00%	0.00%	0.00%

Most Frequent to Least Frequent: e n a i d s o r h m t g þ l w u æ c f y ð b p x z {j, k, g, v, ÿ}

Compare to Modern English: e o t h a s i n r d l u y m w f g c b p k v j q x z {þ, æ, ð, ÿ}


Old English: Single Letter Frequencies
percent
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter


Old English and Modern English
OE
ModE
percent
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter

As suspected, [n] had a much higher frequency in Old English than it does now (9.90% versus 6.31%). The frequencies for [t] and [h] have shot up a considerable amount in Modern English, partly as a result of eth and thorn dropping out from the language. In addition, the lack of the Old English infinitive suffix and system of case marking seems to recruit more [t]s and [o]s to form the preposition so often used in Modern English, namely "to." Notice how the two letters [t] and [o] trend towards similar percentage values:

[t]: 4.04% → 8.50%
[o]: 5.02% → 8.52%

Last but not least, the letter [e] remains most frequent of all the letters. Whether the English in question is several centuries old or freshly new from today, this very special letter has continued and still continues to characterize the English language with its highest frequency usage.

V. Time Warp to the 14th Century: Middle English

i. Linguistic Structure in the Mix

There will be less time devoted to Middle English structure, since, as one could imagine, its linguistic properties fall somewhere *in between* Old English and Modern English. Historically, a lot of shifting was taking place in the grammar that led to more diverse styles of writing out the language. For example, I used Mark Davies's *Polyglot Bible* to count letter frequencies in Luke and the *Wycliffe Bible* to count frequencies in the book of Psalms. The Middle English I viewed for Luke is categorized as "English - 1300s." The Wycliffe Bible was published in 1395. Since translators always have some influence on the text they rewrite and because dialectal variance is inevitable (especially during the years Middle English was used), it is understandable that the writing systems would vary from one another.

While counting letter frequencies for Middle English, I noticed that [ð] and [ÿ] show up, to some extent, in the book of Luke, but not at all for any of the Psalms. Despite this difference, this letter "conflict" genuinely portrays the constant change in language structure and the influence people have on its linguistic behavior.

As stated previously, [æ] and [ð] most definitely dropped off from Old English and were not used in the writing system of Middle English. [þ] remained a part of Middle English, but the letter cluster [th] started to become more common in representing the sounds that thorn and eth once stood for. As for new additions to the alphabet, [j], [k], and [ÿ] began to take on some form in the language, though very minimal. Note that [ÿ] represented a vowel sound, close to the sound one would produce when saying "feet," except that it is articulated with the tongue farther back in the vocal tract.

Another change that took place is found in the inflectional morphology. Recall the [-ian] and [-an] suffixes that were used for Old English infinitives. In Middle English, [-en] becomes the infinitive marker. As for noun inflections, plural and possessive forms were commonly suffixed with [-es], whereas in Old English, these forms could be indicated by [-as] in certain contexts and [-es] in others.

Keeping all this in mind, we predict that the frequencies for [t] and [h] move up from their Old English values, but should still yield values lower than those for Modern English. The absence of [æ] may cause the usage of other vowels to increase; perhaps [a] is a good candidate for this notion. And the frequency of [e] should go up because of the [-en] infinitive marker. I'll let the data take it form here.

ii. Letter Frequencies

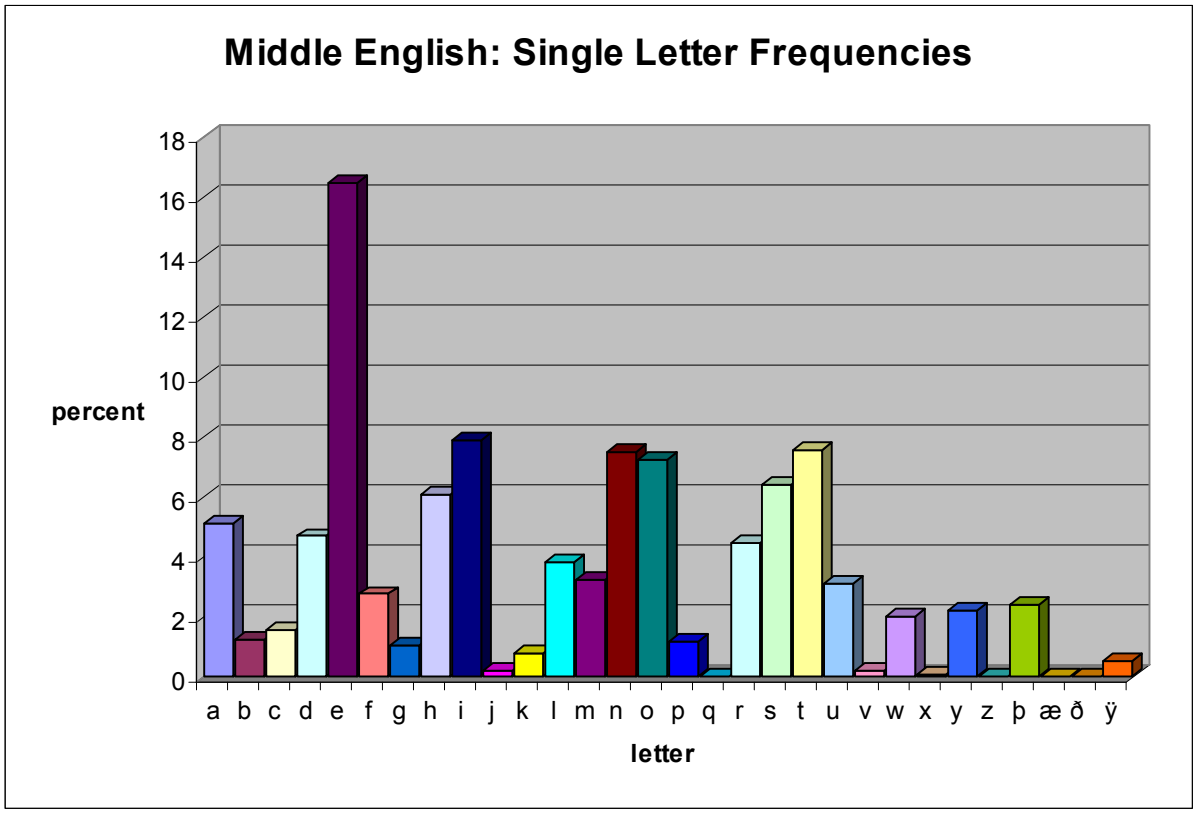
ME	Luke 5	Luke 9	Luke 14	Luke 24	Psalm 63	Psalm 107	Psalm 139	Psalm 143
total	3470	5400	2895	3725	891	3078	1716	1042
a	5.10%	4.76%	5.60%	4.72%	4.71%	5.33%	5.94%	4.89%
b	1.64%	1.04%	2.18%	1.58%	1.12%	0.75%	0.99%	0.77%
c	1.41%	1.28%	1.45%	0.94%	1.68%	1.75%	2.04%	1.82%
d	4.09%	4.63%	3.97%	4.64%	4.15%	6.24%	5.36%	4.61%
e	17.90%	19.28%	17.51%	19.41%	14.03%	16.76%	13.11%	13.82%
f	2.56%	2.74%	2.63%	2.68%	3.14%	2.99%	3.03%	2.40%
g	1.07%	1.00%	1.28%	1.15%	1.23%	1.17%	0.93%	0.67%
h	3.60%	4.56%	4.18%	4.03%	8.19%	8.09%	7.98%	8.25%
i	7.32%	7.31%	7.67%	7.81%	8.87%	7.89%	8.10%	8.16%
j	0.46%	0.39%	0.06%	0.24%	0.34%	0.09%	0.00%	0.19%
k	0.98%	0.59%	0.41%	0.46%	1.12%	0.75%	1.05%	0.77%
l	3.78%	3.52%	3.01%	2.93%	4.71%	3.77%	4.08%	4.89%
m	2.77%	3.61%	2.94%	2.87%	3.14%	2.44%	3.73%	4.41%
n	8.76%	6.89%	7.74%	8.70%	6.51%	8.38%	7.34%	5.57%
o	7.44%	7.67%	7.70%	7.17%	6.62%	6.66%	7.05%	7.58%
p	1.44%	1.26%	1.38%	1.13%	1.46%	0.91%	0.99%	0.77%
q	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.12%	0.19%
r	3.37%	4.20%	4.49%	4.56%	3.82%	5.52%	4.72%	5.09%
s	7.96%	7.69%	6.70%	6.52%	6.29%	5.46%	5.89%	4.70%
t	5.39%	5.37%	7.29%	5.48%	9.54%	8.28%	8.92%	10.46%
u	2.77%	2.83%	2.90%	2.23%	3.37%	2.66%	3.61%	4.70%
v	0.26%	0.24%	0.07%	0.30%	0.45%	0.16%	0.35%	0.00%
w	2.45%	2.00%	1.31%	2.68%	1.91%	2.27%	1.57%	1.82%
x	0.00%	0.05%	0.17%	0.03%	0.22%	0.03%	0.05%	0.00%
y	1.64%	1.61%	1.31%	1.53%	3.37%	1.62%	3.03%	3.45%
z	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
þ	4.87%	4.54%	5.01%	4.81%	0.00%	0.00%	0.00%	0.00%
æ	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ð	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ÿ	0.98%	0.94%	1.00%	1.40%	0.00%	0.00%	0.00%	0.00%

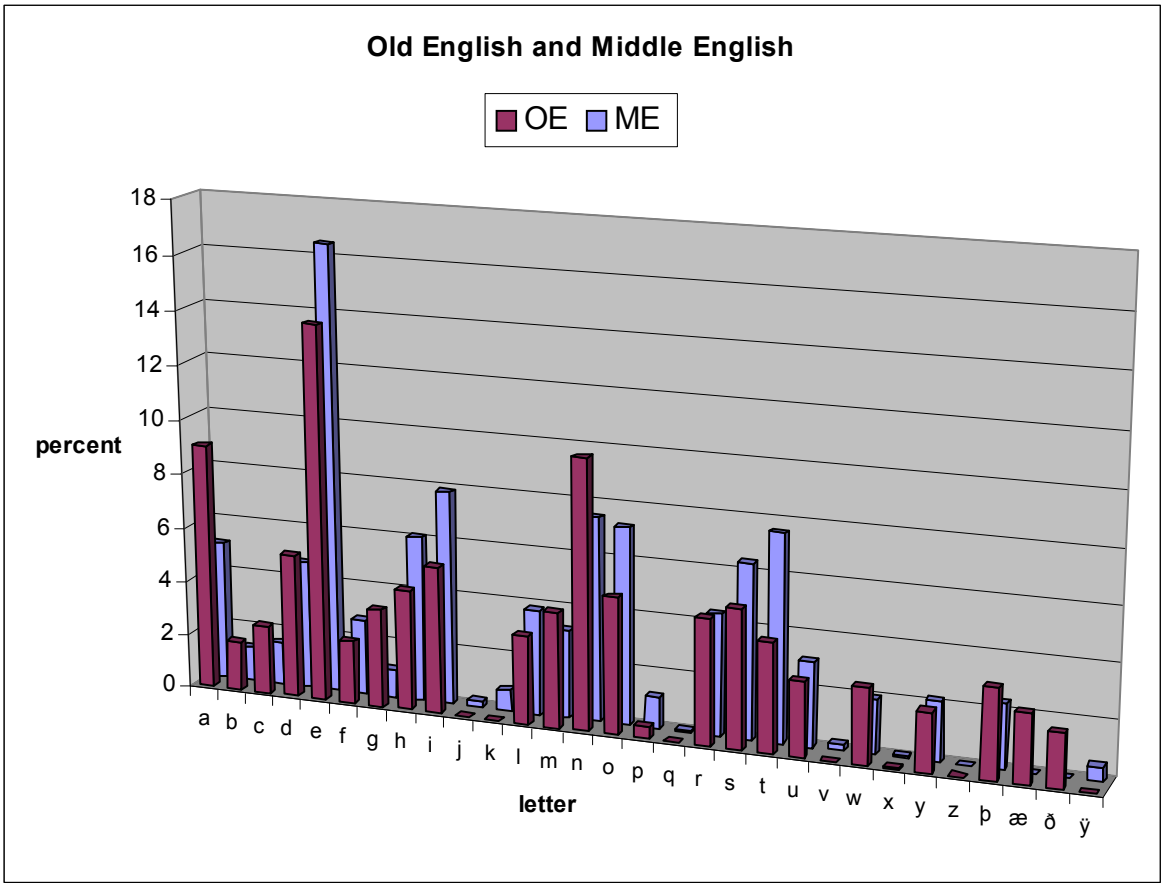
ME	Luke avg	Psalm avg	Overall avg
total	3872.5	1681.75	2777.125
a	5.05%	5.22%	5.13%
b	1.61%	0.91%	1.26%
c	1.27%	1.82%	1.55%
d	4.33%	5.09%	4.71%
e	18.53%	14.43%	16.48%
f	2.65%	2.89%	2.77%
g	1.13%	1.00%	1.06%
h	4.09%	8.13%	6.11%
i	7.53%	8.26%	7.89%
j	0.29%	0.16%	0.22%
k	0.61%	0.92%	0.77%
l	3.31%	4.36%	3.84%
m	3.05%	3.43%	3.24%
n	8.02%	6.95%	7.49%
o	7.50%	6.98%	7.24%
p	1.30%	1.03%	1.17%
q	0.01%	0.08%	0.04%
r	4.16%	4.79%	4.47%
s	7.22%	5.59%	6.40%
t	5.88%	9.30%	7.59%
u	2.68%	3.59%	3.13%
v	0.22%	0.24%	0.23%
w	2.11%	1.89%	2.00%
x	0.06%	0.08%	0.07%
y	1.52%	2.87%	2.20%
z	0.00%	0.00%	0.00%
þ	4.81%	0.00%	2.40%
æ	0.00%	0.00%	0.00%
ð	0.00%	0.00%	0.00%
ÿ	1.08%	0.00%	0.54%

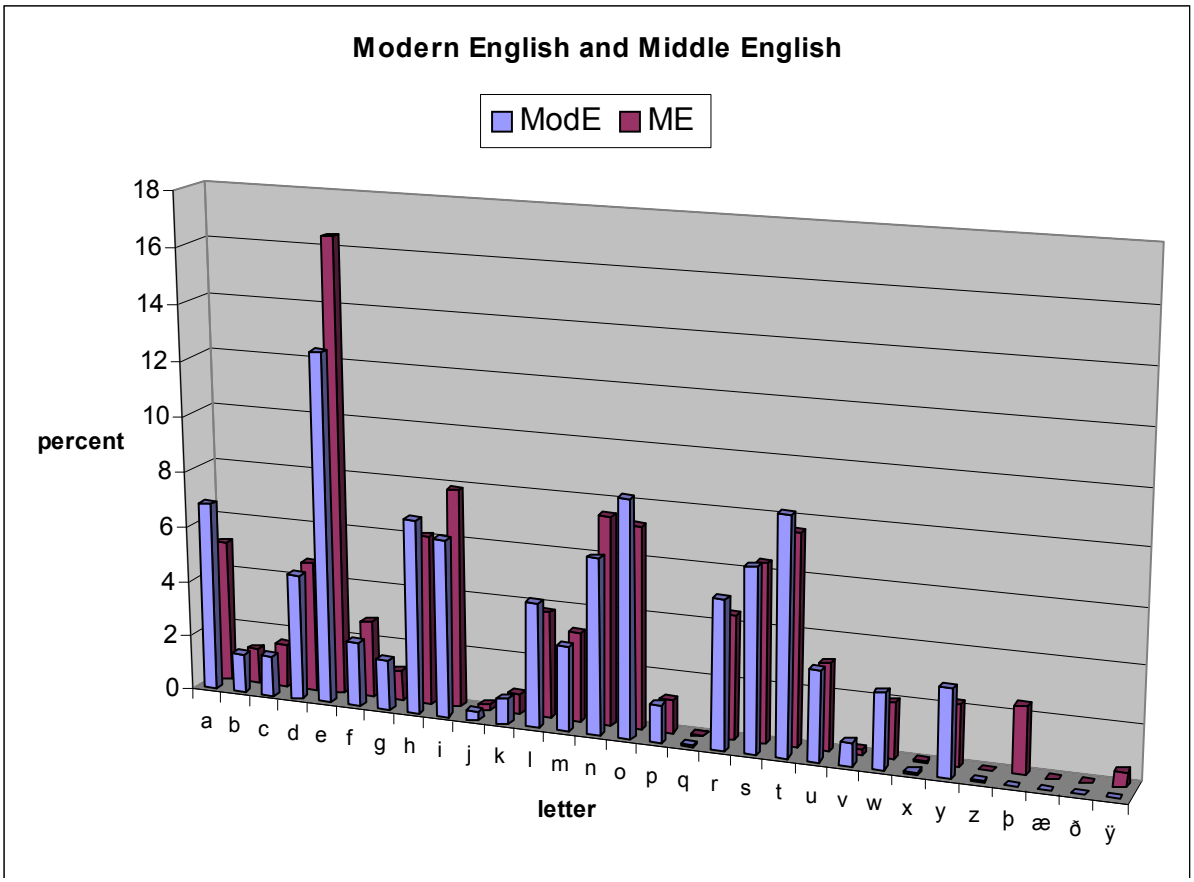
Most Frequent to Least Frequent: e i t n o s h a d r l m u f þ y w c b p k ÿ v j x q z {z, æ, ð}

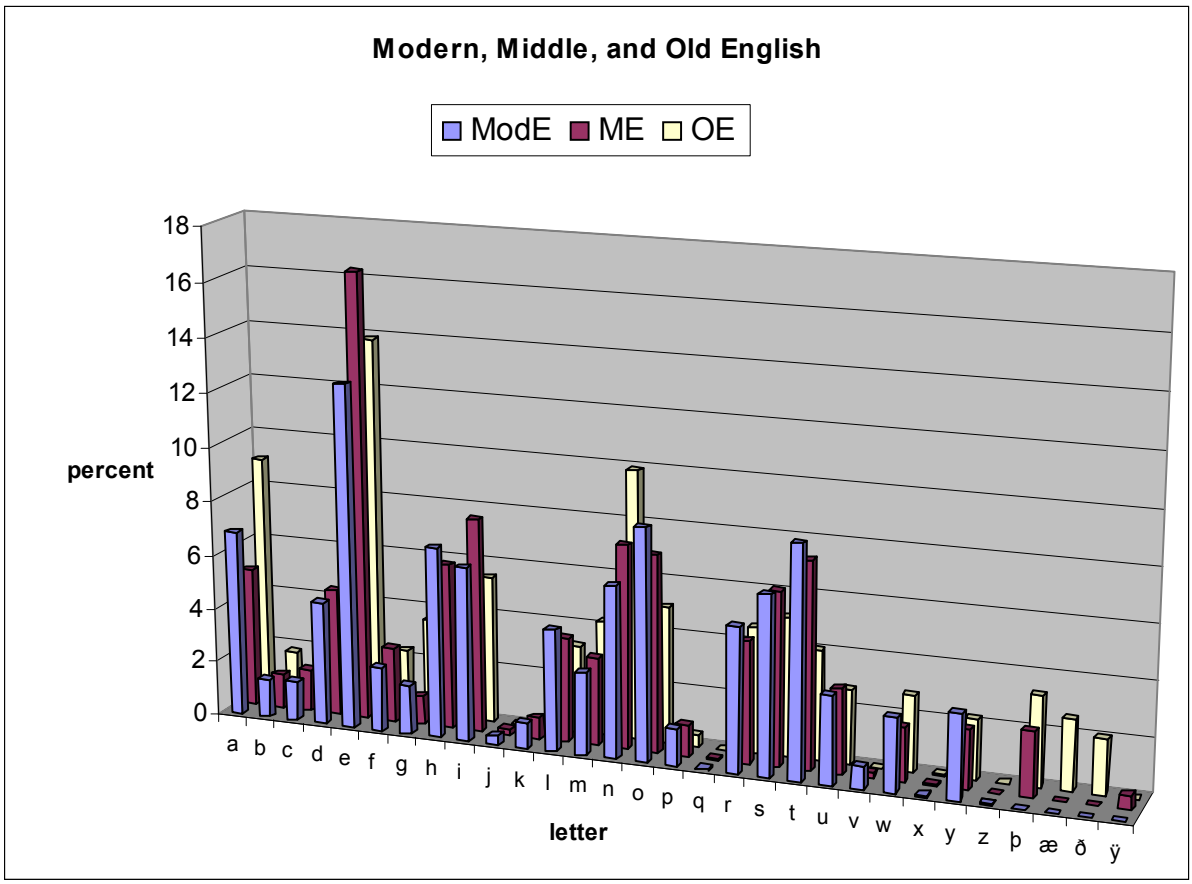
Compare with Old English: e n a i d s o r h m t g þ l w u æ c f y ð b p x z {j, k, g, v, ÿ}

Compare to Modern English: e o t h a s i n r d l u y m w f g c b p k v j q x z {þ, æ, ð, ÿ}


Middle English: Single Letter Frequencies
percent
18
16
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter


Old English and Middle English
OE
ME
percent
18
16
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter


Modern English and Middle English
ModE
ME
percent
18
16
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter


Modern, Middle, and Old English
ModE
ME
OE
percent
18
16
14
12
10
8
6
4
2
0
a b c d e f g h i j k l m n o p q r s t u v w x y z þ æ ð ÿ
letter

These charts and graphs support my hypotheses for letters [t], [h], and [e] -- that the frequency of each would be greater than its Old English counterpart. The frequencies for [t] and [h], indeed, move up in Middle English, but they do not exceed their corresponding values, exhibited by the Modern English data. [e] skyrockets to 16.48%, a value much higher than the 13.82% of Old English *and* the 12.67% of Modern English, yet still holding its title as the most frequent letter.

All these findings are in line with the changes made in Middle English grammar, but we run into one minor problem in the data. I had proposed that [a]'s frequency might increase since [æ] no longer existed in the orthography of Middle English. What we find is the exact opposite; [a] is lowest for Middle English, highest for Old English. I cannot offer a well-defined explanation for this. The vowel represented by [ÿ] may have alleviated [a], showing up in environments which would have otherwise been occupied by [a], had [ÿ] not existed in the language. To better understand this riddle, I would suggest finding lexical items -- ones containing [ÿ] -- and then comparing them to their counterparts in Old English and Modern English. By inspection, it may be possible to detect which vowels in our current spelling system correspond to [æ], and if [a] was ever one of them. Even with the possibility that [ÿ] occurred in contexts similar to [a], that only tacks on an additional 0.54% to the original value of [a], 5.13%. In light of this matter, we should also recall the changed infinitive marker for Middle English; it went from [-ian]/ [-an] to [-en]. So that is likely to contribute to the drop in [a]'s usage. Even though the predictions for [a] are not entirely satisfactory, we were right in saying that vowels in general would increase for Middle English frequencies. Besides [a], all the vowels had increased frequency to some degree.

VI. Findings

On the basis of Biblical texts, all of which embody the English language through time, I was able to compare the letter frequencies of Modern English, Middle English, and Old English. Tracing back to 11th century English, we find a myriad of structural properties that no longer manifest in the language we speak today. Moving forward in the time spectrum, we see the helter-skelter situation for Middle English as pieces of Old English grammar began to fade and as glimpses of new language constructions were set into view.

In general, there seems to be a strong correlation between the increased frequencies for [t] and [h] and the omission of [ð] and [þ]. The change in infinitive markers, as well as the loss of case marking inflection, caused some letters to move up and others to move down from their previous state. For example, the infinitive suffixes for Old English, [-ian] or [-an], brought on a high frequency for [a] and [n], but when the infinitive suffix became [-en] during the days of Middle English, we notice that [e]'s frequency goes up a great deal and [a]'s frequency significantly moves down. And finally, the infinitive marker of today, the preposition "to," certainly shows up in the data, right around 8% for both of these letters (see Modern English: Single Letter Frequencies), which is higher than their frequencies calculated for Old English and Middle English.

A fascinating asset to this study is that, of all the letters ever incorporated in English orthography, [e] *was* and still *is* the most frequently used letter in the language. So no matter what time in history we refer to, we are correct -- and truthful -- in saying that [e] is the most common letter in English. Our findings verify that this is so, and chances are, it will stay that way (given that one thousand years has not changed this unique trait of English).

VII. Conclusion

From the various sets of data, we can easily see how Old English diverges from Modern English. Of course, Middle English serves as some medial point between the two; it still, however, shows very different results from Old and Modern English in terms of frequency analysis. Linguistics reveals how different structures of syntax and morphology make Old English, Middle English, and Modern English very distinct from one another. Frequency analysis further evidences the fact that all three versions of English are different and even deserving to be treated as and considered as *separate* languages.

Single letter frequencies only tell so much about a language. Should cryptographers ever share interest in this topic, they would find that double-letter frequencies in Old English are much different from the ones we see now. Some quick examples are that “god” and “god” (both from Old English) represented “god” and “good,” respectively. The difference shows not in the spelling but, rather, is heard in pronunciation. “God“ (*good* in Old English), was pronounced like how we say “god” (*deity*) today except that the vowel within the root was held out longer. Long vowels and short vowels were contrastive sounds back in the days of Old English, but this was not always reflected in the spelling. So many of the vowel-vowel, double-letter frequencies we’ve examined in Modern English may not be present in Old English. Consider also the consonant clusters in each of the following Old English and Modern English word pairs:

e**cg**/ e**dg**e **sc**ip/ **sh**ip æ**sc**/ a**sh** ni**ht**/ ni**ght** **cn**i**ht**/ **kn**i**ght**

It appears that the consonant clusters used in our current spelling system were represented much differently in Old English. Thus, double-letter frequencies for Old English should not trend in the same direction as today’s double-letter frequencies.

This is just a taste of all the Old English properties yet to be examined from a cryptanalytic perspective. But more importantly, these linguistic shifts in English make us aware of the fact that languages are constantly changing. English will continue to change, as will its letter frequencies. Right at this moment, we are losing gender distinction in the language! Talking about a single person, we often say "they" rather than "he or she," "their" instead of "his or her." A prediction we may draw here is that single letters, [t] and [h], will be used more frequently and their occurrence as an adjacent letter-pair will also increase in English.

Therefore, it is necessary that cryptographers, as well as authors of cryptography textbooks, stay updated with the language. One thousand years from now, "EAT ON IRS" may be of no aid to us when we perform frequency analysis on encrypted text. All languages, prone to innovation, are being linguistically reanalyzed time after time. Under the same circumstances should cryptographers be prepared for linguistic shifts, ready and eager to reevaluate letter frequencies. Let this be a word of advice: expect change and stay updated! □

Sources

1. "The Polyglot Bible." <http://davies-linguistics.byu.edu/polyglot/>.
2. "Old English Pages: Texts and MSS." <http://www.georgetown.edu/faculty/ballc/oe/oe-texts.html#OE%20corpus>
3. "StudyLight.org - Plug in, turn on, and be enlightened." <http://www.studylight.org/>
4. "Old English language - Wikipedia, free encyclopedia." <http://en.wikipedia.org/wiki/Old_English>
5. "Online Bible Translations." <http://www.geocities.com/onlinebibletranslations/>
6. "BibleGateway - New King James Version." <http://bible.gospelcom.net/versions/index.php?action=getVersionInfo&vid=50>
7. "Decrypting text - code breaking software." <http://www.richkni.co.uk/php/crypta/freq.php>
8. "Count Letter Frequencies in Text." <http://klein.math.okstate.edu/PARI/cgi-bin/FrequencyCount.cgi>
9. "English 419: A Chapter on Old English." <http://asstudents.unco.edu/faculty/tbredehoft/UNCclasses/ENG419/OE419.html>
10. "English 419:A Chapter on Middle English." <http://asstudents.unco.edu/faculty/tbredehoft/UNCclasses/ENG419/ME419.html>