Spelling Unfamiliar Names

S. Fitt

Centre for Speech Technology Research

University of Edinburgh

Introduction

This paper will examine the written transcription of unfamiliar spoken names. It is well documented that the writing of personal and place names by people who are unfamiliar with the spelling of the name contributes to the evolution of names. The current paper describes a study which examines the processes involved, using experiments in which Scottish subjects are asked to write down unfamiliar spoken British and European town names.

Writing down unfamiliar spoken names, whether native or foreign, causes a number of problems. If, for example, making a map of an uncharted area, the written form may be based on the spoken form alone (though see Nicolaisen (1961) for a more in-depth approach to transcribing previously unrecorded place-names[1]). This problem can also occur when writing down people's names or addresses. Of course, the writer can often ask for the spelling, but if transcribing from, say, a recorded message this is not possible.

English has a notoriously variable relationship between pronunciation and spelling, so an unknown spoken name may be transcribed with a number of different spellings. Furthermore, there are a large number of names and name-elements which have more than one accepted orthography, so even familiar names can cause problems - we may know that a person's name is [ɹid], but not

whether it is <Read>, <Reid> or some other variation.[2] If the name is foreign, it may contain sounds which have no obvious orthographic equivalent in English. Additionally, folk-etymology can play a part, adapting the unfamiliar to the familiar.

Mishearing is another difficulty, and with foreign-language names there is the further problem of non-native sounds, leading to either accurate perception followed by an attempt at spelling using either native or non-native graphemes, or perceptual categorisation in terms of native sounds, followed by a native-type spelling. We will see below that people do sometimes use non-native graphemes, and they may also use non-native sound-to-spelling correspondences.

Experiments

In order to see how people perform this task, experiments were designed, creating a controlled situation which reduced the number of variables that occur in the natural process. Sixty town names from six different countries were recorded onto tape by a Scottish phonetician, who produced the names as closely as possible to the pronunciations in each language of origin, with the British towns having Scottish pronunciations. Well-known towns were avoided, as were well-known name elements, such as *-land* or *-berg*. Twenty-seven subjects from Edinburgh, aged 14-16, were asked to write down each town name after hearing it twice, and also to choose the country of origin from a closed set of six (Britain, France, Germany, Greece, Italy and Norway).

Responses matching original orthography

In this study it is perhaps irrelevant to talk of 'correct' responses, since for a given spoken name there may be several perfectly legitimate spellings, but only one which is 'correct'.[3] In some cases there may even be more than one existing spelling, for instance Greek <kh> is often transliterated as <ch>. However, it is

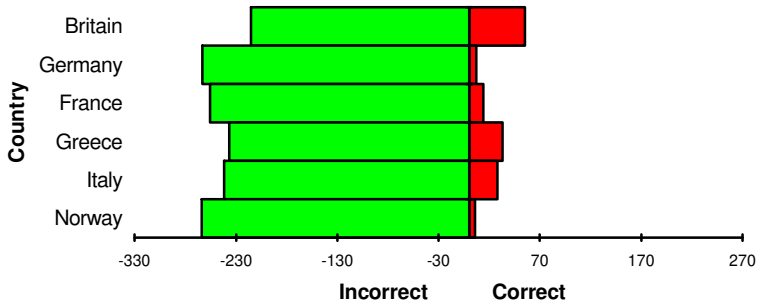

Figure 1: Summary of responses matching original orthography

interesting to see the pattern of matching responses (see Figure 1). British towns have the greatest number of matching responses, but this is not especially high. Although most subjects knew either French or German, while none knew any of the other languages in the study, there were fewer 'correct' responses for France and Germany than for Greece and Italy. We could speculate that the relatively good performance on Greek and Italian towns is due to a simpler vowel system, or to sound-to-spelling correspondences which match more closely the most common ones for English ([a] → <a> and so on), but further data would be necessary to investigate this. It should also be remembered that the towns were not selected randomly; with a random selection of unknown names, the score for British towns in particular would doubtless have been much higher as familiar morphemes such as *-field* would have appeared in the name set.

Legitimate spelling variation

In English, any orthographic vowel may represent schwa, though some do so more commonly than others. Sixteen schwas were present in the prompts, giving 432 responses. Thirteen represented original orthographic <e>, while 3 names had <o> in the original spelling. <e> was in fact the most common response (199) followed by <a> (69) and <i> (66). It is difficult to say whether the responses reflect a general correspondence of [ə] → <e>, in preference to [ə] → <a> and so on, as data of this kind is difficult to obtain. Schwa is particularly problematic, since many words with schwa have variants with either full vowels (such as *obey*) or syllabic consonants (such as *chasm*), so to determine the

statistical likelihood of schwa representing <e> would require extensive speech data, rather than dictionary citations.

For other vowel sounds too there are numerous different orthographic representations (see Venezky 1970).[4] Spelling often varies according to position in the word, for example *Psakhna* ([psaxˈna], Greece). Both the vowels in the prompt were the same, yet whereas the first was unanimously transcribed <a>, while the second had 21 <a>'s but also 5 <ah>'s and 1 <as>. This was possibly an attempt to represent [a] in an open syllable, which does not occur in English, though the rest of the data does not show a clear pattern. (Final orthographic <a> is not especially uncommon in English words, though a large proportion of these are names such as *Clara*, suffixes such as *-phobia* or borrowed words such as *ikebana*. It might be worth examining whether people are aware of such differences between names and other words.)

Consonants may also have legitimate spelling variations. For instance, *Pfinztal*, *Velen* and so on were given single <l> by some subjects and double <ll> by others. Explanations for some responses are more complex, such as *Meyssac* ([mɛ.sak], France), which was given 17 <s>'s and 10 <ss>'s. French does not have lexical stress, but if subjects heard the first syllable as stressed they should have written the [s] as <ss> (as in *lesser*), since intervocalic single <s> following a stressed short vowel is generally pronounced as [z], as in *closet*; however, if they thought the second syllable was stressed (a common interpretation of French words) the single <s> would be a valid spelling for [s], as in *aside*.

Folk-etymology

There are a number of examples of known morphemes being used to transcribe the names. For example, *Bredgar* ([ˈbɹɛd.gaɹ], Britain) was given the spelling <Bred-> by 14 subjects, but <Bread-> by 8 subjects. Of course, it is difficult to say whether the subjects were spelling this word by direct analogy with the word *bread*, or by the use of spelling rules gleaned from a wide variety

of words, which would allow [ɛ] → <e> (commonly), or [ɛ] → <ea> (less commonly).[5] However, for 14 [ɛ] prompts, giving 378 responses, there were 292 <e>'s while the only <ea>'s were the 8 responses for *Bredgar*, suggesting analogy with *bread*. It should be noted, though, that <ea> is used for [ɛ] disproportionately often in the environment preceding [d]. (Bredgar had the only [ɛd] sequence in the data.)

A number of instances of <-shire> appeared in the responses. *Bolkesjö* ([ˈbɔl.kə.ʃø], Norway) was given 6 <shire>'s despite containing a non-English vowel and no final [ɹ], as would appear in a Scots pronunciation of *-shire*. *Sollom* ([ˈsɔ.lʸəm], Britain) was, unsurprisingly, spelt <Solemn> by 6 subjects. Strangely, other subjects appear to have taken elements of *solemn*, with 4 using <-umn>, though final <mn> is relatively rare in English.

It can also be the case that people try to apply their knowledge of foreign languages in processing unknown foreign names. For example *Livorno* ([liˈvoɾ.no], Italy), was written by one subject as <Les Vorno> and by another as <Les Vernos>, and placed in France; the subjects were evidently using their knowledge of French to interpret the name.

Categorisation of foreign sounds

> "The phonological system of a language is like a sieve through which everything that is said passes...when [a person] hears another language spoken he intuitively uses the familiar "phonological sieve" of his mother tongue to analyse what has been said."[6]

A good example of the problem of categorising spoken foreign sounds is *Tallard* ([ta.laːʁ], France). The French /t/ is part-way between English [t] and [d]; 10 subjects wrote <t>, while 17 wrote <d>. Sometimes there is an obvious native counterpart to a non-native sound, such as [ç] in *Ekhinos* ([ɛˈçi.nɔs̲], Greece), which was mostly given similar spellings to Scots [x]:

Spelling	Occurrences
ch	12
kh	3
h	9
th	2
kih	1

Table 1: Spelling of [ç] in Ekhinos

Of course, it is not possible to tell whether the subjects perceived the sound as [x] (or in some cases [h]), or whether they perceived it correctly as [ç] and used the most appropriate spelling they could.

Mishearings vs misperceptions

Mishearing involves a major error in hearing a sound, while misperception describes the erroneous categorisation of a sound in terms of the native system, and is a possible explanation for *Ekhinos*, as described above. If a sound is misperceived, some of the original features are preserved: for instance, the French vowel [y] contains the features [+front] and [+rounded]. It is typically nativised by English speakers by changing one of these features, resulting the high back rounded vowel [u], or sometimes split into [ju], thus preserving all the features but distributing them across two phones.

It therefore seems likely that the [ɣ] of *Megara* [ˈmɛ.ɣa.ɾa] was generally misheard, rather than misperceived, as the most common spelling given was <n>, whose phonetic equivalents bear no resemblance to [ɣ]. Some sounds are more liable to be misheard than others, due to their acoustic qualities.[7] Also, some prompts were misheard more often than others, because of the quality of recording, unclear pronunciation and so on.

In some instances it is clear that sounds were simply not heard, as a large number of subjects omitted to transcribe any letter at all for a particular sound; sometimes graphemes were inserted where there was no corresponding sound, either through mishearing or an error in writing. An interesting problem arises from the use of post-vocalic <r>. Given that the subjects were Scots, with rhotic

accents, that the prompts were spoken by a Scottish speaker, and that the foreign languages in the study mostly use graphemic <r> to represent an [r] sound of some description, we would not expect subjects to use the spelling <r> unless they actually hear an [r].[8] However, there are a number of instances in which subjects did in fact write <r> where there is none in the prompt, for example *Snåsa* ([ˈsnɔˠ.sɑ], Norway), which had 10 <r>'s written after the final vowel. It is possible that the subjects are influenced by RP sound-to-spelling correspondences; they may draw on RP because in a formal environment such as an experiment they use their knowledge of standard English pronunciation, or because they are treating RP as a 'foreign language'.

Use of non-native graphemic features

Some non-native graphemes and grapheme sequences were used, as well as non-native sound-to-spelling correspondences. For example, two subjects used <ß> for the final sound of *Tsamandas* ([t͡sa.manˈdaʂ], Greece), one placing it in Germany and one in Norway. Additionally, accented characters such as <é> were used. Although the majority of these appeared in names which the subjects thought were French, there were a few in towns which subjects placed in other countries, which contradicts the usual view of nativisation that involves only the source and borrowing languages.

An example of a non-native sound-to-spelling correspondence is <Sch-> in *Schapen* ([ˈʃaː.pʰən], Germany), and *Schwenke* ([ˈʃvɛŋ.kʰə], also Germany). Sixteen subjects did in fact use <Sch-> *Schwenke*, and 2 for *Schapen.* The discrepancy between the two is perhaps due to the perception of the name; 20 subjects placed *Schwenke* in Germany, possibly due to the stereotypical German [ʃv], while only 5 did so for *Schapen.*

Representation of length and rhythm

A number of words had long vowels or long consonants in the prompts. Geminate consonants are not typically found in the middle of monomorphemic English words (though they may be found in polymorphemic words, such as *bookcase*). In some cases it is not possible to tell whether the subjects perceived the long consonants, since a word-medial double-consonant spelling such as <kk> can represent a single spoken consonant in English. However, some double consonant spellings must represent two sounds, for example word-medial <pn>. Looking at the data, we find five names with phonologically long consonants (see Table 2).

	[bː] in *Bobbio* (Italy)		[pː] in *Copparo* (Italy)		[kː] in *Dokka* (Norway)		[lː] in *Hellesylt* (Norway		[ŋː] in *Lyngen* (Norway)	
	Data	**No.**	**Data**	**No.**	**Data**	**No.**	**Data**	**No.**	**Data**	**No.**
Probably long	mp bp lp rp	3 1 3 1	mp np rp	6 1 1	nk nc	5 8			gn	2
Ambiguous			pp	1	kk	1	ll	10		
Probably short	p b	9 10	p	17	c ch ck	7 1 5	l r	14 1	g ng	20 5
Other			blank	1			blank	2		

Table 2: Representation of long consonants

Some spellings suggest that subjects have heard extra length; as the table shows, this often manifests itself as a continuant preceding the consonant. Unfortunately there is little data on short consonants in similar environments for comparison.

Conclusions

This experiment produced complex data, some of which gives clear indications of the way subjects processed the names, and some of which can be interpreted in a number of ways. We can see that the subjects are not linguistically naive; although they sometimes interpret unknown names, both

foreign and native, using their native language framework, they also employ their knowledge of foreign languages, sometimes overgeneralising this knowledge to languages they do not know. Further work is now needed in order to build up a model of the interaction of the many processes involved.

Notes

1 Wilhelm F.H. Nicolaisen, 'The Collection and Transcription of Scottish Place-Names', *Septieme Congresso Internazionale de Scienze Onomastiche,* (Firenze, 1961, Vol. 4), 105-114.

2 Angle brackets are used throughout this paper to represent orthographic forms; pronunciations are given in a close IPA transcription, in order to differentiate between the different languages involved.

3 Original spellings listed here were taken from the *The Times Atlas of the World, Concise Edition*, (London, Times Books, 1992, 6th edn.).

4 Richard L. Venezky, *The Structure of English Orthography.* (Paris, Mouton, 1970).

5 The role of analogy in pronouncing words is discussed in Robert J. Glushko, 'Principles for Pronouncing Print: the Psychology of Phonology', In Alan M. Lesgold and Charles A. Perfetti (eds), *Interactive Processes in Reading,* (Hillsdale, Erlbaum, 1981), 61-84.

6 Nikolai S. Trubetzkoy, *Principles of Phonology,* (Berkeley, University of California Press, 1939 [translation 1969]), 51-2.

7 George A. Miller and Patricia E. Nicely, 'An Analysis of Perceptual Confusions among some English Consonants', *Journal of the Acoustical Society of America,* (Vol. 27, 1955), 338-52.

8 French *-er* verbs, of course, are a counter-example.