

Surname Variation and Surname Matching Algorithms

INITIAL DRAFT

Peter Christian (*peter@spub.co.uk*)

6th May 2002

0 Conventions

In line with normal linguistic practice angled brackets indicate spellings, slashes indicate pronunciation, e.g. initial <wr> corresponds to /r/ in English. However, I have sometimes been vague, as it has not always been clear to me whether it would better to work with (messier) spellings or (more precise) sounds.

1 Preliminaries

1.1 *Why are there variants?*

There are a number main reasons for the spelling variation in English surnames.

- 1) The fundamental reason is that English has a "deep" orthography (compared to, for example, Spanish or German) — many rules are required to relate a spelling to the pronunciation it is associated with. English spelling is archaic and preserves many features lost in the spoken language, but with names there are often written forms which reflect the contemporary pronunciation.
- 2) Complex grapheme↔phoneme rules. Because an English phoneme is often represented by a number of graphemes, a writer who does not know an existing spelling for a name will have a number of options for representing it orthographically. Conversely, and in contrast to the present-day conventions for the normal vocabulary, a distinction of spelling does not mean a distinction of referent.
- 3) Speech style. Different styles of speech on the scale formal↔casual mean that there will be a variety of pronunciations for a single name, even within a single community. Again, spellings reflecting various pronunciations of a single surname give rise to variant forms, some of which may become stable written forms in their own right (rather than ad hoc variants).
- 4) Dialect variation. Different dialects have different rule sets (e.g. whether /r/ after a vowel is pronounced as a consonant or modifies the preceding vowel). Where spellings reflect output rather than input, different regions will show different spellings. This is a minor issue with consonants but is one of the reasons why vowel matching is problematic.
- 5) Affixation. A number of affixes are found in British names, producing forms which often interchanged with the affix-less form, most notably:
 - a. the genitive <-s>, which is either a patronymic or a locative suffix;
 - b. patronymic affixes derived from words for "son"
 - c. locative prefixes such as "at/a", or "de" in French-derived names.

Other minor causes of *instances* of variation, rather than stable variants, are:

- 6) Hypercorrection. A minor source of variants in written records are hypercorrect forms, e.g. the Hadams and Hadamson in the 1881 census are the result either of an

enumerator correcting the speaker's presumed dropped /h/ or of the speaker trying to give what he thinks is the more formal version of his name.

- 7) Mishearings/mispronunciations. In fact, these will generally be indistinguishable from some of the variation arising from the causes listed above. Potential mishearings are likely to produce spelt forms which resemble properly constituted English names, and will often be indistinguishable from genuine variation. Many supposed errors of this type will not be the result of acoustic problems (speaker drunk, toothless, etc.) but will in fact be *misinterpretations* of what is heard (see Hypercorrection, above). The extreme case will be non-English names where the writer may have little idea how to interpret what he has heard, and no suitable conventions for representing it.
- 8) Mis-spellings. What this means is far from clear, because surnames have never been subject to the same standard of "correctness" that has developed for the normal vocabulary. However, mis-spellings can be identified where they produce forms which could not be regarded as with the structural rules for English words (for example, the 8 names beginning <bn> in the 1881 census). We should expect these to be small in number, and the most obvious ones will often be from the modern transcriber.

One final issue:

- 9) Spelling pronunciation. Literate speakers have a tendency to pronounce names as they are spelt (feeling this to be less "sloppy" than authentic pronunciations they may hear). However, since this does not generate new spellings, it can be ignored from the point of view of ToBS.

In all we can establish 3 types of variation that a good Thesaurus must capture

- Orthographical variants — Variant spellings which are neutral with respect to pronunciation e.g. <Brown> / <Browne>.
- Phonological variants — Variant spellings which reflect differences between standard spelling and pronunciation, or between different pronunciations.
- Morphological variants — Those with and without various patronymic or locative affixes (Mac, de).

1.2 Types of Phonological Variation

The main types of what one might call phonologically motivated spelling variation are:

1. substitution, i.e. one sound replaces a similar one, or one grapheme replaces another which can represent the a similar sound (e.g. Fidler<>Vidler, Stephens<>Stevens)
2. deletion, i.e. a sound is lost (typically an unstressed vowel, one consonant from a cluster, a consonant which is vocalised) and the spelling reflects this.
3. insertion, i.e. addition of a sound, reflected in spelling, usually as a glide between two others, e.g. Thomson -> Thompson.

The effects of 2) and 3) are in fact very similar in the variant pairs they produce, and can be treated as a single phenomenon. Most purely orthographical variations can be treated as pseudo-phonological variations of the same sort, not least because that's how they started out.

1.3 Morphological variants

The relevant suffixes are well known and their number is small. For the approach proposed here, it would probably be preferable to identify these in pre-processing and possibly strip them for matching.

2 Towards a better matching process

There are two basic sources of the inadequacies of Soundex-like name matching algorithms.

- 1) The binary nature of their matching technique: two surnames either have the same code or they don't;
- 2) Their primitive phonological knowledge, particularly
 - a. lack of context-sensitivity
 - b. over-clumping of consonants
 - c. inability to deal with vowels

Phonex, of course, shows that it is quite possible to make advances in 2), specifically 2a), but 1) remains a fundamental limitation on the successfulness of the approach.

Of course, these algorithms were designed to be simply expressed and not processor-intensive, and from that point of view these two "failings" should not really be regarded as such.

However, for our project, we can work on the premises that:

- 1) any amount of pre-processing can be permitted since it is not carried out at run-time but rather the results stored in the thesaurus for subsequent use in matching
- 2) multiple pre-processing results could be stored for a single name, allowing for different criteria or different weightings
- 3) the processing power of current technology permits more demanding matching algorithms
- 4) knowledge of grapheme-phoneme relations has improved, and algorithms which encapsulate that knowledge (usually as part of text-to-speech systems) can be drawn on
- 5) a large corpus of data (particularly complete national datasets like the 1881 census) will allow some scope for statistical approaches to matching, possibly combined with geographical information.

These mean that both limitations of Soundex-style systems are no longer design requirements.

An additional benefit is the limitation of the scope of our project to British surnames, which means that it can be optimised for particular patterns of variation attested in British documents.

The approach I am going to outline in this paper decomposes the phonological structure of surnames, so that potential variants can be checked to see how their structures are mappable. At this stage, I have considered only consonants; vowel remain problematic though some statistical techniques suggested here may be appropriate.

3 An Approach to Consonant Matching

3.1 General principles

I list here some general principles which underlie consonant spellings in English. The first of these is a stipulation in advance of our discussion on what constitutes variant, and is therefore open to debate; the rest are commonplaces of English or general linguistics.

1. Two surnames are variants if they can be mapped on to each other on a component by component basis using standard orthographical, phonological, and (to a lesser extent) morphological rules of English.
2. All syllables have the structure: (onset) + nucleus + (coda)
3. Onset and coda each consist of one or more consonants; the nucleus consists of a vowel
4. There is a characteristic set of permissible ("legal") onsets and codas for English, a subset of which is applicable to surnames (i.e. excluding those found only in foreign words, e.g. the /sf/ in <sphere>).
5. Medial consonant clusters consist of a legal coda followed by a legal onset
6. For onsets/codas with more than one consonant, there are restrictions on permissible combinations and sequences. (Equivalent restrictions exist for other languages, and those for most other European languages differ only in minor details from English.) See 3.1 and 3.2 below.
7. These restrictions are historically *very* stable, and can certainly be regarded as immutable for the period since 1500; dialect differences, if any, will be *very* minor.

It follows from this that:

8. Any sequence of consonants in a written (unabbreviated) name in a primary source should be regarded as an *attempt* to represent a permissible onset, coda, or coda+onset (even if mistakes are made in the execution, or unusual conventions are used).

Illegal clusters found in any dataset will be:

- mistakes
- foreign spellings
- marginal spellings

For example, the 1881 census has 128 names beginning with the illegal onset <sr>: many are manifestly modern transcription errors (SREVENSON, SRIFFITHS); a few are obviously foreign (SRODZINSKI); in other cases <sr> is being used for <shr>, or <str> (SRUBSOLE, SRONG). All are clearly marginal, two thirds of them appearing only once, only 3 appearing more than 10 times. It may be possible to map the mistakes and marginal spellings to "acceptable" spellings, though it seems better to exclude them for initial work.

3.2 Orthographical problems

Before analysing consonant clusters there are some orthographical issues that would need to be tackled in a pre-processing stage. These are not necessarily trivial!

1. Deciding when an orthographical single vowel (usually <e>) is not a nucleus is a necessary preliminary to correct identification of medial consonant clusters (e.g. to show that Sidgwick and Sidgewick are variants). I'm not sure this can be formulated in an algorithm, but there will certainly be heuristics available.
2. How to decide where the stress should be is not obvious to me, but again there are undoubtedly heuristics if not algorithms for doing so in the normal vocabulary. How far these might be applicable to surnames remains to be seen.
3. Deciding when medial <sh>, <th> represent one sound or two (e.g. Westham) will also need an appropriate technique.
4. In the case of spelt consonant combinations where one consonant is not pronounced (e.g. <wr>, <kn>) it would be simpler to reduce these in pre-processing. However, care would need to be taken that e.g. Plowright and Pinkney are not reduced to Ploright and Pinney. Spellings like <gh> which sometimes correspond to no spoken sound present a similar problem.
5. Where spelt consonants correspond to multiple pronunciations, one would generally need to look at neighbouring sounds, so there would be some need to evaluate vowel spellings to allow <c> to be interpreted as /k/ or /s/. (Alternatively, one could allow both and evaluate the results statistically at some later stage.)
6. There is a good case for treating names with initial vowels as starting with a zero consonant. This simplifies matching names with and without <h>, <y> and perhaps <w>.

3.3 Initial consonant clusters (onsets)

The following table sets out the permissible *spoken* consonant clusters in English names. These correspond to about 60 different spellings in all.

1 st slot	2 nd slot	3 rd slot					
sibilants	stops & fricatives	glides					NORMAL SPELLINGS
		r	l	m	n	w	
optional	mandatory	optional (max. 1 of)					
s	0		+				l, sl
				+			s, m, sm
					+		c n, kn, sn
						+	w, wh, sw
S		+				sh, shr,	
z						z	
s	p	+	+				p, pr, pl, sp, spr, spl
	b	+	+				b, br, bl
	f	+	+				f, ph, fr, fl, (phr, phl rare)
	v						v
s	t	+				+	t, tr, tw, st, str (stw not found)
	d	+				(+)	d, dr, (dw rare)
	T	+				+	th, thr, thw
s	k	+	+			+	k/c, kr/cr/chr, kl/cl, qu, quh (Scots) sk/sc/sch, scr, squ
	g	+	+			(+)	g, gr, gl, (gw rare)
	h						h
	j						y
	C						ch
	J						j (d - before /u/ only)

Notes

1. I've used Sampa conventions (<http://www.phon.ucl.ac.uk/home/sampa/english.htm>) for sounds where the IPA uses a symbol not available in ASCII:
S= sh T=th C=tch J=dj
2. The table has been designed from a pragmatic point of view, so some liberties have been taken with the presentation of the patterns from a theoretical point of view.
3. Strictly, /j/ should be in a column next to /w/, but is only found before /u:/ and is not in any case represented in spelling.
4. The boxes in slot 2 group sounds which are typically regarded as interchangeable in Soundex-like systems.
5. The table excludes clusters found only in words/names from other languages (e.g. /sf/, /Sl/), though it is worth noting that names from other European languages could be catered for without altering the basic structure of the table.
6. The specific issues of Welsh and Scots names have not been addressed, but these would only require additional entries in the table, not a change in its structure.

3.4 Final consonant clusters (codas)

Coda rules are more difficult to analyse, but from a pragmatic point view these two tables are more or less satisfactory.

1a liquids		1b nasal		2	3 stops & fricatives	4 t suffix	5 s suffix
max. 1 of		max. 1 of					
<i>r</i>	<i>l</i>	<i>m</i>	<i>n</i>	<i>s</i>		<i>t</i>	<i>s/z</i>
(+)	+	+	+				+

1a/b				2	3	4	5
max. 3 of							
max. 1 of							
<i>r</i>	<i>l</i>	<i>m</i>	<i>n</i>	<i>s</i>			
(+)	+	+		+	p	+	+
(+)	+				b		+
(+)	+	+			f	+	+
(+)	+				v		+
(+)	+		+	+	t		+
(+)	+		+		d		+
(+)	+		+		T/D		+
(+)	+		+	+	k	+	+
			+		g		+
(+)	+		+		S		
(+)	+		+		C		
(+)	+		+		J		

Notes

1. This table is more complex than Table 1. and is less comprehensive — there are some low frequency combinations which cannot be represented by this structure (e.g. names ending in /kst/ <xt>). There are also some combinations here which are not permitted, but I've left for the sake of clarity (/mspts/, for example). I will add a list of individual spellings at a later stage.
2. The two possible positions for <s> are required because a coda can have two /s/ components. However, the final <s> is always morphological and could be treated separately, relieving this table of a complication.
3. <r> is present in spelling, but absent from pronunciation of many dialects, including the standard, or rather it combines with the preceding vowel.
4. There are a number of orthographical consonant clusters which are not represented in this table, e.g. <mb>, <ght>, <bt> because either (a) only one of the spelt consonants corresponds to a spoken consonant, or (b) they belong to separate syllable and are therefore not a true coda (compare <lamb> and <lumber>).
5. Words ending in <le> after a consonant have a vocalic /l/, which can be regarded as identical to /el/.

3.5 Medial clusters

In principle, medial clusters can only be constructed from a legal coda followed by a legal onset. In practise there are only a fraction of the 4,000 or so combinations of the components given in the foregoing tables, and many will be quite rare.

3.6 Matching clusters: an example

To look at knowledge of consonant structure can be used to provide improved surname matching, I have taken a group of names from the FamilySearch data from the PRs of the two neighbouring parishes Pevensey and Westham, East Sussex, which seem likely to represent variant spellings for a single family.

RAYNOLD, RAYNOLE, REINOLDS, RENNALS, RENNOLS, RENOLDS,
REYNOLDS, REYNOULD

If we ignore the vowels and double consonants (as we would at this stage) we can reduce these to four basic types, distinguished by the final consonant cluster

		coda					
		1a	1b	2	3	4	5
a	Raynole	L					
b	Rennols	L					S
c	Raynold	L			D		
d	Reynolds	L			D		S

Here is a set of variants that would not be matched by Soundex (R540, R545, R543), and yet patently constitute a group of related forms.

For all these forms, the structure of the whole name is identical up to the L, and thereafter two slots are at issue. So we then need to ask, for each of these almost-matching pairs, whether the difference between them can be expressed in a rule which is valid in English pronunciation or word-formation.

a::b	L::LS	yes, genitive/patronymic -s
a::c	L::LD	yes, consonant cluster simplification
a::d	L::LDS	don't know, 2 differences in the same coda, but it could be a strong form of consonant cluster simplification if there are frequent examples on other names and words
b::c	D::S	no, different slots
b::d	LS::LDS	yes, consonant cluster simplification
c::d	LD::LDS	yes, genitive/patronymic -s

The four established pairings would then remove any doubts about the other two on a second pass by implication: if (a) is a reduced form of (b) and (b) is a reduced form of (d), then (a) is a plausible variant of (d); (b) and (c) are both variants of (d) and therefore of each other.

3.7 Issues

The "consonant cluster simplification" would not be simply a descriptive phrase — we'd have to be sure that the simplification of LDS to LS, say, is exemplified in English generally. The whole case can be evaluated by looking at how many other possible instances of this rule might be found in the data (for example, SHIELD, and RUMBOLD show the same pattern of variation in the 1881 census).

It is worth noting that many rules are directional, and it is not simply a matter of interchangeability. In the example in 3.6, LS can be derived from LDS in a hard form, i.e. if there is a surname ending in LDS, then a form in LS *will* be a legitimate variant of it. But LDS cannot be derived with the same certainty from LS — for example LFS might also simplify to LS, but we would not want that to imply that names ending in LFS and LDS are interchangeable.

This example has dealt with consonant insertion/deletion. But a look up table would also improve the matching of consonant substitutions. For example Soundex treats PBFV as if they are all equally interchangeable, but this is far from true. It would be *very rare* to see a name with -V- appear in a spelling with -P- (at best a French name with V, where a Latinate form with P is also found) but V/F interchange is much more frequent. The obvious rule-of-thumb to adopt is that the more distinctive features sounds have in common, the more likely they are to be interchanged. F differs from V by one feature (voice), P from V by two (voicing, manner of articulation). Likewise two sounds that occupy different slots in the cluster structure cannot be interchanged.

3.8 Automating the comparisons

We would not need to make such detailed comparisons for every pair of forms individually. It would make sense to construct look-up tables to contain such rules, possibly with some weighting as to their frequency or plausibility.

This approach would mean 3 tables (onsets, codas, coda+onsets) with cells containing information on the plausibility of regarding the components on the two axes as interchangeable. If this is done for each cluster in a pair of names, then we can develop an overall measure of consonantal similarity. Note that this separation of the matching criteria from the algorithm will allow improvements to be made to the matching without altering the algorithm itself.

There are obvious run-time implications for comparing every pair of the 400,000 names in the 1881 census (though it's only around 75,000 if you ignore spellings with a frequency of less than 10). But with this technique, at the first non-match, one would simply abandon a comparison, so there would be no need to compare the 10,000-odd forms beginning B+vowel with the 8,000 or so beginning L+vowel, because no matter what their subsequent similarities they cannot be variants. And obviously the tree would be pruned at every comparison.

3.9 What's missing?

Unfortunately it would be more problematic to match vowels by the same process, not least because dialect variation will be much more significant, and the grapheme↔phoneme correspondences are much more complex.

However, if one started from the basis of a set of names with compatible consonant structures, one can envisage a variety of ways of estimating which vowels spellings are most readily interchangeable with which others in particular context. So there is no problem in principle with developing equivalent vowel matching tables. But that is a quite separate task.

Also, of course, there is interaction between neighbouring consonants and vowels.

But a consonant matching algorithm of the type described would still represent a significant improvement over Soundex-like systems in accuracy of matching, even without any vowel component. The consonant matching would clump maximal groups of possible variants, which vowel matching would be able to reduce. Particularly for monosyllabic names, a consonant-only system offers poor results.