Beider-Morse Phonetic Matching

Alexander Beider & Stephen P. Morse
(September, 2008)

Background

Searching for names in large databases containing spelling variations has always been a problem. A solution to the problem was proposed by Robert Russell in 1912 when he patented the first soundex system. A variation of Russell's work, called the American Soundex Code, was used by the Census Bureau to facilitate name searches in the census. A major improvement to soundex occurred in 1985 with the development of Daitch Mokotoff (DM) Soundex by Randy Daitch and Gary Mokotoff. DM Soundex is a soundex system optimized for Eastern European names. Both of these soundex systems have, nevertheless, a major disadvantage – they generate many false hits, requiring the researcher to wade through a lot of extraneous matches.

The phonetic-matching method, developed by Alexander Beider and Stephen Morse, attempts to alleviate that situation. The initial work on this algorithm was based on the article by Alexander Beider "Some Issues in Ashkenazic Name Searches" (*Avotaynu: The International Review of Jewish Genealogy*. Vol. XXIII, Number 1, Spring 2007, pp.3–13) and the long term desire of Stephen P. Morse to improve the engine of his various online searchable databases (*http://stevemorse.org*) including Ellis Island Passenger Lists.

The initiation of this project (and, more precisely, the personal meeting of its two authors in Newark in July 2007 and their decision to work together) was made possible by the sponsoring provided by the International Institute for Jewish Genealogy and Paul Jacobi Center in Jerusalem and the organisational efforts by Sallyann Amdur Sack and. The two authors would also like to thank Logan Kleinwaks, Gary Mokotoff and Jean-Pierre Stroweis, who tested the draft versions of BMPM and provided numerous valuable comments.

Main Principles

The main objective of Beider-Morse Phonetic Matching (BMPM) consists in recognizing that two words written in a different way actually can be phonetically equivalent, that is, they both can sound alike. But unlike soundex methods, the "sounds-alike" test is based not only on the spelling, but on linguistic properties of various languages.

For common nouns, adjectives, adverbs and verbs this task is of limited interest. Except for orthographic and typographic errors, these words rarely have spelling variations. The situation is different for proper nouns (i.e., names) – they can appear in documents written in different languages and spelled according to the phonetic rules of the language of the document. Determining that two different spellings correspond to the same name becomes even more difficult when the two spellings use letters from different alphabets.

In its current implementation, BMPM is primarily concerned with matching surnames of Ashkenazic Jews. This is due to the list of languages whose graphic and phonetic features are already taken into account. The name matching is also applicable to non-Jewish surnames from the countries in which those languages are spoken. However the structure of BMPM is general and we are already planning to extend it to additional languages so as to make it applicable to the surnames of Sephardic Jews, as well as non-Jewish names from the corresponding countries.

BMPM is designed to be used as a programming tool, and an individual would be very hard-pressed to do the calculations manually. To use the system, a user would enter a name on a form, that name would be transmitted to a server running the phonetic engine that would generate the BMPM values, and those values would then be compared to the BMPM values that were previously generated for all the names in a specific database.

Overview of the Algorithm

The spelling of a name can include some letters or letter combinations that allow the language to be determined. The current version of BMPM includes about 200 rules for determining the language. Some of them are general, while others include the context in which they are applicable (e.g., beginning or the end of a word, following or preceding certain letters). The processing of these rules yields one or several languages that could, in principle, be responsible for the spelling entered by the user.

One option of the BMPM engine allows for specifying the language explicitly. That would apply when the database is known to be in a specific language, in which case each name in that database can be encoded using the rules of that language, and the language-determination test need not be done.

In a number of languages, forms of surnames used by women are different from those used by men. So once the language has been determined, rules for that specific language are applied that replace feminine endings with the masculine counterparts.
After the name has been de-feminized, the phonetic engine tries to identify the *exact phonetic value* of all letters of the name, and transcribe them into a phonetic alphabet. This, too, is language dependent, so the rules that perform this phonetic transcribing are also language-specific. If it was not possible to uniquely determine the language, the phonetic engine processes the name using generic rules.

Once the name is processed by either the language-specific rules or the generic rules, the phonetic engine applies to the resulting string of phonetic characters a series of phonetic rules that are common to many languages. An example is the rule known in linguistic literature as *final devoicing,* which states that at the end of the word the voiced consonants are pronounced as their unvoiced counterparts. Another rule, also applied by the phonetic engine, is that of *regressive assimilation*, whereby a consonant acquires characteristics of the consonant that follows it.

After applying the steps outlined above, the original surname is transformed into a string of phonetic symbols, which we call its *exact phonetic value*. We next apply a series of rules that (1) take into account the fact that some sounds can be interchangeable in some specific contexts; and (2) allow for phonetic proximity of a pair of sounds resulting in their partial confusion. At the end of this step, the initial surname is transformed by the phonetic engine into what we call the *approximate phonetic value*.

An additional step can be applied for names that were originally written in Hebrew characters and transliterated into Latin or Cyrillic characters. Depending on the preferences of the person doing the transliteration, vowels can appear or not at certain positions, and certain consonants can have a choice of phonetic values. As a consequence, a transliteration of the same name from Hebrew to Latin characters made by different people can yield different results. So a series of additional rules can be applied that allow for the ambiguity of certain sounds when dealing with the Hebrew spelling. After doing so, the original surname is transformed into what we call its *Hebrew phonetic value*.

Searching for Matches

Applications of name matching involve searching for names in electronic lists. The phonetic values (*exact*, *approximate*, *Hebrew*) of the name being searched for needs to be generated by the phonetic engine at the time the search is performed. But prior to doing any searches, the phonetic value of each of the names in the list needs to be calculated. Some simplifications can be used when processing the entire list of names because there might be information known about the language and the spellings used within the list.

The matching of individual name to names present in specific electronic lists will result in either an *exact match*, an *approximate match*, or a *Hebrew match* depending on whether a match was obtained with the *exact phonetic values*, the *approximate phonetic values*, or the *Hebrew phonetic values*. The *Hebrew match* is significant only when the name being sought was originally written in Hebrew or, alternatively, the name in the electronic list was originally spelled in Hebrew characters. If the user knows that neither situation applies, then the *Hebrew match* is of no importance and can be simply ignored.

Comparison to Daitch-Mokotoff Soundex

As mentioned above, soundex is one of the solutions proposed in the past to solve the problems of name matching. It has several variants of which the Daitch-Mokotoff (DM) method is the one that is the most commonly used in the domain of Jewish Ashkenazi genealogy.

In soundexing, every letter either receives a numerical value, or is simply omitted. Different consonants can receive the same numerical values. All vowels are treated as interchangeable. In contrast with BMPM, soundexing does not search for the equivalence of sounds: even different (but sometimes close) sounds can match. Consequently, when matching names, soundexing may have a significantly larger number of false positives than BMPM. On the other hand, it can find some true matches that are not found by BMPM because the equivalence is not purely phonetic.

The domain in which soundex seems to be more appropriate than BMPM is when the original European form of the name (which is the form as it appears in the list) is not known and all that is known is the Anglicized form of the name as used today.

There are other contexts in which BMPM is more appropriate than DM. These include

(1) Automatic processing by computer of large databases in order to find matches between elements of various databases,
(2) Searching for individual original names (not yet anglicized) in large databases.
(3) A special group of matches identifiable by BMPM that are not found by the current version of DM Soundex.

In brief, BMPM greatly advances techniques for name searching and identification. At the same time, BMPM and DM act complementary tools. Each of them has contexts in which its application is more appropriate than the other or that of any other method.

*****