IBM Information Management software

Names that Sound Alike, but are Not Spelled Alike: The Use of Phonological Information in Automatic Name Searching

Richard Lutz
Stephan Greene

Contents

1. Introduction

The field of computational linguistics has matured and expanded as the power and speed of computers has increased, memory and storage costs have fallen and authoring languages when increased in capabilities and level of sophistication. As a result, information extraction and retrieval techniques have also made remarkable progress. In the "Information Age", there are simply too many data to be analyzed and sorted manually. With ever increasing accuracy, algorithms are processing and extracting relevant information from a wide variety of sources and from data in an increasing number of languages.

One data type that has been persistently problematic for automatic processing is that of named entities, especially personal names and names of organizations. Unlike other data elements, such as Social Security numbers or other kinds of IDs, named entities can show significant, sanctioned variation. Use of nicknames versus formal given names, use of initial versus full spelling of a name, presence or absence of maiden names, and presence or absence of titles or qualifiers (such as Dr. or Jr.) are just some of the more obvious ways that a name as label can vary. Search engines constantly confront such issues of form and format.

Furthermore, names tend to be much more variable in spelling than other lexical items. Even the most frequent American surnames can have many common spellings (e.g., Connely, Connelly, Conley, Conelly). Names transferred from non-Roman script into Roman show variation as well (e.g., Qadafi, Khaddafi, Ghadafi, Khadaffi, etc.). Predicting the way a particular name of a particular individual will be spelled is often problematic. Success in searching for a name, particularly if the actual spelling is unknown, can be extremely haphazard, as anyone acquainted with Internet search engines can verify. Effective name search engines are badly needed in areas such as the airline and hotel industries, telephone directory assistance, the health professions and governmental agencies-in short, anywhere that large databases of names are maintained and searched.

This paper describes the results of a government sponsored research project that investigated the utility and feasibility of incorporating phonological information about names into the process of automatic name searching. The goal was to determine whether including information about the pronunciation of names in the search algorithm could improve upon search methods based exclusively on spellings, that is character-based comparisons. The strategy taken was to try to predict probable pronunciations of names based on language-specific orthographic rule sets, and then automatically measure the phonological similarity between the query name and potential matches in the database based on those pronunciations. Retrieved names could then be returned in ranked order, with names most like the query name ranked towards the top of the list. The result of this project is a working prototype that accepts single input names (e.g., "Smith" or "John"), searches a preprocessed database of names, and returns names that vary in spelling but show some phonological similarity with the input name (e.g., "Smythe" and "Schmidt"; "Jon" "Jan", "Gianni"and "Joan").

2. Statement of Problem

It is reasonable to wonder why personal names seem to challenge automated matching and retrieval systems and why a ranked list of phonologically similar names might ever be useful. It is unusual to think that a person's name poses any sort of general difficulty. After all, a name belongs to a person; that person "knows" his/her name and uses that name for personal identity. There is an assumed association of a name and a single person and therefore personal names are viewed as fixed items, much like numbers.

However, the apparently inseparable link between the name and the person can be broken when a name is entered into a database. There is now a dissociation of the name and the individual, so the ability of the name to discriminate uniquely is reduced, if not eliminated. The name in the database may refer to many people with the same name or a similar name; the name now selects a group of individuals. Clearly, when a database contains more than one John Smith, additional information must be used to distinguish one record from another, so that only relevant records are considered.

There is an additional and more confounding problem with names in large databases: the variability associated with how names are entered and stored.

How a name is stored within data records may, and often does, deviate in form from the way it is entered at the time of query. Indeed, personal names pose special problems in terms of data retrieval because names exhibit much more variation in form than do other lexical items. As with all lexical items, part of this variation is error-based and random:

for instance, a name might be mistyped (e.g., Jpnes). Much of the variation, however, falls within a well-understood and predictable set of parameters associated specifically with names. A nickname may be substituted for a formal given name, for example, or the Hispanic matronymic name might be entered in the "last name" field, when the patronymic would be more in line with what an Anglocentric culture means by "last name". A field might be left blank entirely (e.g., middle name). Of more immediate concern to this paper is the orthographic variation inherent in names as a data type.

The word chair can refer to any members of the set of chairs, but its written form is fixed by standard English orthographic conventions. Names such as Leigh or Johansen, Stephen or Jeffrey have a number of common spellings, and probably a number of uncommon ones as well.

Because of the dissociation created between label and referent when a name is entered into a database, however, a spelling mismatch between a name being queried and the relevant name or names in a database may occur. The circumstances for this mismatch are common, and include:

- oral transmission of a name (e.g., to a telephone operator);
- guesses on the part of the person entering the name (e.g., an Internet search for a name heard on the radio: Wooster entered for Worcester);
- changes over time or across cultures (e.g., genealogical histories: Vulchansky for Vlèansky; Beecham for Beauchamp); and
- different ways of transcribing from non-Roman script into Roman (e.g., Xie, Hsieh, Sye, etc.)

Automatic name searching may frequently benefit from algorithms that can account for variation in a principled way, and that can retrieve fuzzy matches of names that are somehow similar in pronunciation to the query name.

Character-based name searching algorithms rely on spelling as the basis for calculating distance between the query name and the database name. While spelling using Roman characters is not unrelated to pronunciation, the relationship between the two is often inconsistent, and the orthographic information (i.e., conventions of the spelling system of a language) is at times misleading. Thus, one spelling may map to multiple pronunciations: Lutz can be pronounced to rhyme with puts, cuts or shoots, and at least several additional non-English pronunciations are possible. The converse, of course, is also the case: there may be a number of ways of representing a single pronunciation: Lewis and Louis, for example, are commonly pronounced identically by English speakers.

Character-matching techniques assume a reliable relationship between the orthographic system and the pronunciation. This assumption is flawed because the goodness of fit between orthography and pronunciation, especially for English, is many-to-many; that is, a given Roman character can stand for more than one sound, and an individual sound may be represented in more than one way in the spelling system. Thus, the sound [f] can be written as f (Frank), ff (Taffy), ph (Phillip) or even gh (Rough). Conversely, the gh digraph may represent the [f] sound of Rough, be silent (Dough), or represent

[k] (in some pronunciations of McClaughlin), [h] (in Monagham), [g] (in McGhee) or [gh] (across syllable breaks, as in Bighouse).

Orthography, of course, is language-specific. Thus, while the letter X may generally stand for [z], [ks] or [gz] in English, it represents a [dz] in Albanian and an alveopalatal fricative in the Pinyin transcription of Mandarin Chinese. Thus predictions about pronunciations of names must be based on language-specific orthographic conventions.

Historically, name searching techniques have relied either on character-based comparisons (e.g., n-grams that calculate the percentage of paired letters shared between two spellings), or on key-generated sets (e.g., the Soundex system of classification, where differences between letters such as c, s, g, j, k, q, x and z are leveled). Both approaches are problematic, because comparisons based on standard spellings may mask phonetic similarity. Thus, the name Knox shares few letters with the name Nocks even though they are generally pronounced the same. The approach taken in the current project is to process names by a series of rewrite rules that transcribe the spellings into likely strings in International Phonetic Alphabet (IPA) transcription. A comparison of the IPA representations of Knox and Nocks will reveal an exact match.

Furthermore, once names are converted into IPA notation, phonologically-based comparisons become possible. A name such as Knox might then be compared to names with similar but not identical pronunciations, such as Nock or Noggs. When the exact name is not known or unclear, a phonologically-based search engine becomes an attractive retrieval tool.

The following describes the approach taken in the current prototype.

1. Names in Roman characters are automatically processed by a statistically-based algorithm that predicts the likely cultural source of the name based on character patterns. Currently, names are analyzed as Arabic, Mandarin Chinese, Hispanic or "Other".

2. Based on the automatic culture classification, names are processed by a set of rewrite rules for the appropriate culture. The rewrite rules automatically convert the Roman spelling into a regular expression with International Phonetic Alphabet (IPA) notation. One spelling of a name is thereby represented in IPA as multiple possible pronunciations.

3. In addition to the culture-specific rewrite rules for Arabic, Mandarin Chinese and Hispanic names, all names are processed by a generic set of rewrite rules that approximate likely pronunciations based on standard conventions of English orthography.

4. Names that share at least one potential pronunciation with the query name are returned as "exact phonetic matches", ranked by spelling. Thus, a query on the name Knox might return Knox, Nocks, Nauckes as exact matches.

5. The IPA representation of the query name is compared with those for the names in the database. For each comparison, a score is returned based on a phonologically-based similarity metric. Names are retrieved and ranked based on the phonological score, as well as on other factors, including number of syllables, likely cultural classification and initial consonant. Key to this project was the development of a definition of similarity that would adapt phonological principles of Optimality Theory in order to retrieve names that varied in spelling from the actual query name, but which were somehow close enough in terms of pronunciation to warrant consideration as a possible match. It is the nature of the similarity metric used in this procedure that is the subject of this paper.

3. Optimality Theory

3.1 Theoretical Foundations

The view of phonology espoused by Optimality Theory (Prince and Smolensky 1993) provides the overarching framework for the development of the phonological similarity metric. Optimality Theory (OT) marks a move to a fully constraint-based view of phonology. The theory argues that a universal set of violable phonological constraints exists in Universal Grammar. Individual phonological competence consists of a specific, established ranking of these constraints. A rich base of possible phonological forms is constrained to surface forms that violate the lowest ranked constraint possible for each form.

Work in OT has established formulations of specific constraints and some specific, universal constraint rankings (sub-hierarchies) that appear to recur within language-specific constraint rankings. These constraint and ranking formulations form the basis upon which the phonological similarity metric is constructed. However, the metric's evaluation procedure necessarily departs from strict observance to the fundamentals of OT. OT is a theory of phonological competence, not a theory of performance. In particular it is not developed as a theory of how to evaluate the similarity of related surface forms that may be the result of the introduction of performance errors. Here, performance errors are construed quite broadly to include not only articulatory, but auditory, typographical, and transliterational errors or inconsistencies. In the literature of generative linguistics, it is argued by Chomsky and others that "it is not incumbent upon a grammar to compute" (Prince and Smolensky 1993). This is assumed in OT, as Prince and Smolensky consider OT to be a formalism that seeks to bridge the gap between high-level symbolic grammatical theory, and lower, near-neural level computational, connectionist (or perhaps stochastic) implementations of grammar. As such, in OT constraint violations are infinite. If constraint C1 dominates constraint C2, no number of violations of C2 can ever overcome a single violation of C1. The power of the mechanism of constraint domination has been well argued. Its framework and general results have proven useful here as well, but in the end, for the purpose or measuring phonological similarity, constraint violation must be quantified in some way as something other than infinite. It is in this sense that the evaluation procedure most significantly departs from the principles of OT.

The OT framework proved to be more facile for adaptation to the problem of measuring phonological similarity than traditional Phonotactics + Repair theories. The notion of constraint domination, while softened from the position of infinite domination, played a central role. For the purposes of the similarity metric, the implementation of the constraint model was more straightforward than a Phonotactics + Repair model would have been. Phonotactics + Repair models require rules, with conditioning environments, to describe phonological changes. There are several problems with this. To begin with, conditioning environments are not available (or practical) in the prototype system's runtime environment. Secondly, typical rule-based grammars do not easily accommodate arbitrarily posited phonological changes. Lastly, even if such rules were devised, and could be applied in a system with conditioning environments available, there is little theoretical foundation for scoring the application of one rule as more of less costly than the application of another rule. Phonotactics + Repair models employ a multiplicity of processes to achieve a singular target. OT, by contrast, employs the single process of ranked constraint violation to achieve a multiplicity of targets. The linguistic principle of markedness provides a foundation for scoring these constraint violations. The concepts of constraint domination and markedness thus provide a straightforward framework within which to devise a cost model for arbitrary phoneme transformations. It is simply a matter of quantifying the constraint violations required to achieve the posited change.

The design of the phonological similarity metric can be seen in some sense as Scalar Phonology in an OT Framework. Scalar Phonology is a theory in which features are multi-valued, rather than binary or privative. The multiple values on a scale can be formalized as an ordinal scale. The Sonority Hierarchy is perhaps the most widely accepted scalar construct in phonology, although formulations deriving it from the primitives of binary or privative theories have likely contributed to its wide acceptance. Gnanadesikan (1997) argues for development of the scalar framework within OT. She proposes the idea of ternary scales as representations that more naturally account for a number of problematic phenomena, such as chain shift. Her work focuses primarily on the Inherent Voicing (IV) Scale. This scale essentially collapses the features [voice] and [sonorant] into a single, ternary scale:

The Inherent Voice Scale (Gnanadesikan 1997):
voiceless obstruents ↔ voiced obstruents ↔ sonorants

In this theory, a segment bears no binary or privative value for voicing or sonorance, but is simply identified as being a voiceless obstruent, a voiced obstruent, or a sonorant. Gnanadesikan cites a good deal of evidence for this scale, including both phonological and phonetic evidence that the steps on the scale between voiceless and voiced obstruents, and voiced obstruents and sonorants, are not in some sense 'equal'. This is based on the idea that there are two kinds of voicing, that of marked voicing in obstruents, and inherent, spontaneous voicing in sonorants. The costs associated with changes in voicing and sonorance in the similarity metric are at least generally congruent with the IV scale. This ternary scale approach thus appears to provide additional theoretical foundation for the approach detailed below, and may inform later iterations of it.

3.2 Adapting OT to the Prototype Name Search System Design

3.2.1 Phonological Similarity of Strings

The degree to which two IPA variant strings are phonologically similar is computed based on measures of phonological similarity between pairs of corresponding IPA characters in the strings. The string-level similarity metric is determined using an "edit distance" algorithm (Hall and Dowling 1980). The algorithm was originally developed to do approximate string matching to compensate for relatively minor spelling or typographical errors.

The basic approach of the algorithm is to determine the cost of editing one string in order for it to become the other. It does so in a manner which optimizes the total number of editing operations (character replacements) required, and this path of least resistance, or "edit distance", is used as the score of the similarity between the two strings. The algorithm iterates over the characters in the two strings. If two characters are the same, the cost is zero (0) and the cumulative score is unchanged. If two characters are different, a cost of one (1) is incurred and is added to the cumulative score.
3.2.2 Phonological Similarity of Phonemes (IPA characters)
In the prototype name search system, the edit distance algorithm

compares pairs of IPA characters, rather than orthographic characters. Additionally, in order to make it more effective as a measure of phonological similarity, the edit distance algorithm was modified to score pairs of different characters gradiently according to their phonological relatedness, rather than simply as identical (0), or different (1). The system makes active use of 40 phonemes (including the null segment), which come primarily from American English but also Mandarin Chinese, Arabic, and Hispanic (various dialects of Spanish). A 40 by 40 matrix of values ranging from 0.0 to 1.0 was prepared to indicate the scores for every possible change of any phoneme into any other phoneme. This Edit Distance Cost Table is read into the system at runtime and used as needed during the ranking phase.

Two versions of the Edit Distance Cost Table were developed. The first version was based on the model of generative phonology proposed in the classic Sound Pattern of English (Chomsky and Halle 1968) and will be referred to as the SPE-based table. The second version of the Edit Distance Cost Table, now in active use, is based on Optimality Theory and will be referred to as the OT-based table.

In the SPE-based table, gradient scores of phonological similarity were calculated using a simple measure of distinctive phonetic feature differences. Using generally the same set of phonetic features as proposed by Chomsky and Halle (1968), scores were determined by calculating the number of features for which the two phonemes in question differed in value, and then dividing by the total number of features. This approach helped demonstrate the utility of using phonologically based gradient scores within the edit distance algorithm, but was problematic because the scores failed to capture certain important aspects of the phonological landscape. For example, the phonemes [p] and [f] differ in two features, [continuant] and [strident]. The phonemes [p] and [è] also differ in two features, [continuant] and a particular treatment of [place]. Thus the edit distance score of both phoneme pairs is the same. However, the two labial phonemes, [p] and [f], participate in a number of phonological alternations cross-linguistically, and can even be allophonic in some languages. No such set of processes or relationships exist (productively, at the very least) between [p] and [è]. In addition, [p] and [f] and likely to be judged as more similar than [p] and [è], particularly within the context of real words, such as similar personal names. Clearly, scores based simply on counts of phonetic feature differences fail to capture the essential relations among phonemes in a manner that is comprehensive and consistent enough for use within the generalized comparison of phonemes in the prototype name search system.

A very different approach was taken in the OT-based table. The idea of markedness constraints is invoked and made to operate on feature geometry (Kenstowicz 1994) rather than on the linear set of features outlined at the inception of generative phonology in SPE. Using these concepts, a new set of measures of similarity among phonemes was developed. The idea of using markedness constraints as formulated in OT was originally inspired by the Universal Place Markedness Hierarchy that has been proposed in the OT literature

(Prince and Smolensky 1993, Lombardi 1998). The overall approach taken is detailed in the next section. As put into practice, it will be seen that some of the constraint formulations appear more like faithfulness constraints than markedness constraints, but they are presented here as they were originally conceived.

The values populating the OT-based table are generated using a series of ranked constraint hierarchies. As noted above, the OT conception of phonology defines grammars in terms of language-specific rankings of universal, violable output constraints. Here, however, constraints are used to build an abstract cost model that gives a measure of the costs associated with bridging the feature differences between any two phonemes. A final distance cost for any two phonemes is the sum of the individual costs for each violated constraint. The constraint rankings determine the relative weight of each constraint in the calculations. Effectively, each series of ranked constraints has been roughly treated as an ordinal scale (Stevens 1946), where movements of differing distances along the scale are charged (possibly non-linear) costs. A side effect of this is that costs generated in this way are not inherently reflexive: the cost of [X] going to [Y] will not necessarily be the same as the cost of [Y] going to [X].

3.2.3 Constraint Usage in Detail

Constraint formulations were developed for root level features, as well as the autosegmental feature spaces for place and manner of articulation, and the larynx (state of the glottis). Some feature treatments are more grounded in theory than others, some are more grounded on empirical facts than others, and some are basically ungrounded in anything other than the goal of proper operation within the prototype system, but all treatments were formulated with the same general schema.

3.2.3.1 Root-level Costs

Features

The root-level features of concern are [consonantal] and [sonorant]. Constraints are formulated as working against a change in a feature. The notation *^[feature] indicates a constraint against a change in a feature (basically equivalent to a faithfulness constraint). Given that changes in the consonantal feature indicate changes between vowels and consonants, while changes in sonorance may or may not imply that change, a *^[consonantal] constraint dominates a *^[sonorant] constraint:

*^[consonantal] >> *^[sonorant]

This indicates that a change in the [consonantal] feature incurs a larger cost than a change in the [sonorant] feature. Within the constraint against changes in the consonantal feature, a large, symmetric cost is incurred for any change at the consonantal level:

high cost	+consonantal → -consonantal
high cost	-consonantal → +consonantal

Within the constraint against changes in sonority, the Sonority Scale is used. The Sonority Scale orders sounds with decreasing sonority as follows:

Vowels > Glides > Liquids > Nasals > Obstruents (Fricatives > Affricates > Stops)

There is virtually no cost for changing between vowels and glides given their close affinity. A higher cost is incurred for changing between any sonorant and any obstruent. Changes among sonorants and obstruents are costed as follows:

low cost	[vowels] ⇔ [glides]
high cost	[sonorant] ⇔ [obstruent]

Epenthesis and Deletion

In the optimization of editing operations by the edit distance algorithm, sometimes a character difference between two strings is treated as the addition or deletion of a whole character, rather than as a comparison of two extant characters. While these operations are not necessarily a reflection of true phonological processes of epenthesis and deletion, these concepts are used to help guide the measure of the cost of such operations within the system. Epenthesis and deletion can often be highly constrained phonological phenomena, but these processes can usefully capture an appreciable set of relations between personal names, such as that between "Conley" and "Connelly." Epenthesized or deleted characters are treated as part of the root-level cost computation, as these are operations at the segmental level.

The costs associated with epenthesis and deletion are treated symmetrically. Lowest cost is associated with the epenthesis or deletion of vowels, while highest cost is associated to the epenthesis or deletion of obstruent consonants. Glides are assigned slightly higher cost than vowels, with nasals and liquids slightly higher than glides. These costs rise as sonority lowers:

Epenthesis and Deletion:

< lower cost			higher cost >
< higher sonority			lower sonority >
[vowels]	[glides]	[liquids and nasals]	[obstruents]

3.2.3.2 Autosegmental Feature Costs

The remaining phonetic features are separated into treatments of the three commonly used feature spaces of feature geometry: place of articulation features, manner of articulation features, and laryngeal features. Initial project research suggested that changes in place features were most salient in judgements of (dis)similarity, followed by manner features, and finally by laryngeal features as least salient. Thus changes in place are assigned the highest edit distance cost, and changes in laryngeal features are assigned the least edit distance cost. In terms of a constraint hierarchy, where violations of the highest ranked constraints receive the highest costs, the relations among these feature classes is expressed as follows:

*^[place] >> *^[manner] >> *^[laryngeal]

Place Features

For place or articulation, the following Place Markedness constraint hierarchy has been proposed in recent OT literature (Prince and Smolensky 1993, Lombardi 1998) as a universal ranking of universal markedness constraints:

*[labial], *[dorsal] >> *[coronal] >> *[pharyngeal]

This ranking claims that labial and dorsal segments, together, are the most marked. Pharyngeal segments are the least marked. Changes in place in a direction going against markedness (becoming more marked), and thus violating higher ranked constraints, bear the most cost. Changes resulting in less-marked segments incur lower cost. In addition, greater distances traversed on the place markedness "scale" incur higher costs than shorter traversals. Thus, the cost model attributes ascending cost values to the following list of possible place feature changes:

least costly	coronal → pharyngeal
	coronal + dorsal → coronal
	coronal → coronal + dorsal
	labial, dorsal → pharyngeal
	labial, dorsal → coronal
	pharyngeal → coronal
	coronal → labial, dorsal
most costly	pharyngeal → labial, dorsal; labial ↔ dorsal

Note that in the feature system as used in the system, some phonemes are both [coronal] and [dorsal]. The Universal Place Markedness hierarchy suggests that coronals that remain coronal while becoming dorsal are nonetheless becoming more marked. Coronals dropping their dorsal feature are becoming less marked. Thus the loss or dorsality among coronals is less costly than the addition of dorsality among coronals, as shown in the placement of the second and third transformations above. However, the costs are lower than that for other transformations involving the dorsal feature, as the changes in question essentially capture processes of palatalization (as well as "de-palatalization"). Incurring relatively low costs for palatalization, a fairly productive process cross-linguistically, and only slightly higher cost for loss of palatal status among coronals, was considered a reasonable approach based on our instincts about the evidence of such processes in personal name data. The Place Markedness constraints do not address changes between labials and dorsals, as the markedness constraints for these features are not ranked relative to each other. High cost was assigned to changes between dorsals and labials based on the large physical distance within the supralaryngeal articulatory space between dorsals and labials.

Manner Features

The following four manner features are utilized in the Edit Distance Cost Table: [continuant], [nasal], [strident] and [lateral]. No inclusive universal constraint hierarchy of any kind is yet, or likely, to be established for the highly discrete manner features. Thus each feature is treated individually.

Assuming that full oral stops are the most marked manner of articulation, the feature [continuant] is analyzed in light of the following markedness hierarchy for descriptive manner features, corresponding to decreased constriction of the vocal tract:

*[stop] >> *[affricate] >> *[fricative] >> *[sonorant]

These constraints make that claim that stops are the most marked segments; the realization of a stop violates the highest ranked constraint. Setting nasals aside, this translates into the following constraint hierarchy on the [continuant] feature:
*[-continuant] >> *[±continuant] >> *[+continuant]

Thus moving in the direction of a positive value for continuance is less costly than moving toward a negative value for continuance:

least costly	-continuant → +continuant, ±continuant
	±continuant → +continuant
	±continuant → -continuant
	+continuant → ±continuant
most costly	+continuant → -continuant

Relatively small, reflexive costs are incurred for any changes in the simple binary features [strident], [lateral] and [nasal]. The features [strident] and [lateral] are important for distinguishing among certain phonemes, and thus changes in value for these features should be noted. But within this model, such changes do not represent significant degrees of phonological dissimilarity that are not primarily accounted for by other features elsewhere in the model. See section 6 for further discussion of nasals.

Laryngeal Features

Aspiration and glottalization are not distinctive in English, are generally not reflected in English orthography, and thus their consideration is not critical within the system. However, a small cost is incurred for changes in value for the aspiration feature, as this feature serves to distinguish among some of the Chinese phonemes in the system. This cost is not required to accommodate any influence on similarity judgements by native English speakers when assessing Chinese names, but rather to make proper system-internal use of the IPA variant representations of Chinese names.

Voicing is the primary laryngeal feature considered. For changes in voicing, little cost is incurred overall, given the rankings of the autosegmetal feature class subhierarchy, but slightly higher cost is assigned to voicing as opposed to de-voicing. This is based on the current theory claiming that voicing is a privative feature, and is marked in obstruents:

lower cost	+voice → -voice
higher cost	-voice → +voice

These costs serve to distinguish between voiced and voiceless alternates among the obstruents. There are costs incurred for any changes between obstruents and sonorants, which may imply changes in voicing, but these are dominated by the higher ranked root-level constraints.

Cost Calculation

A feature table was prepared to itemize the feature values for all phonemes in use in the system. Using this table as input, a utility program was written to evaluate the constraint violations implied, within the model just described, by the pairwise transformation of all phonemes. Costs were calculated for each pair of phonemes and output to a file for input to the system.

4. Results

The OT-based table of edit distance scores is more in line with current phonological theory, more properly reflective of important phonological relations and processes, and thus ranking results are improved within the systemover the SPE-based table.

Approximately a dozen representatives of the system's sponsoring client participated in a day long, formal acceptance test for the system. The participants were both project coordinators and end-users, people for whom name searching is a major portion of their job responsibilities. The formal test battery included a number of personal names known to be problematic for search systems. Each name was accompanied by a set of sample results against which the results returned by the system were compared. The performance of the system was judged to be significantly better than the various other systems with which the testers were familiar, as well as improved over an earlier, interim version of the system. The general response of the acceptance test participants was exceedingly enthusiastic, and the system was approved for adoption.

In addition, several expert native speaker linguists on the the system development team, experienced with issues related to personal name searching, judged the results as improved over the SPE-based table in extensive side-by-side testing.

5. Limitations

The constraint model used to compute phonological similarity, as described above, appeared to work well overall, but had some limitations. These limitations were generally reflected in the need, in several cases, to manually fine-tune certain values in the cost table. Across the board application of the rules that score constraint violations to the phonetic features in the input feature table did not produce workable values in all cases.

For example, after some testing and evaluation, costs for the epenthesis and deletion of the strident coronal fricatives [s] and [z] were lowered to that of epenthesis and deletion of nasals and liquids, rendering them exceptional among the obstruents. Similarly, epenthesis and deletion of the glottals [h] and [ ], were adjusted to the same cost as epenthesis and deletion of the glides. Thought not terribly surprising, these changes were driven purely by the empirical results generated during testing of the system. The final scores for epenthesis and deletion fall on a scale as follows:

Epenthesis and Deletion:
< lower cost higher cost >
< higher sonority lower sonority >
[vowels] [glides, [h], [ ] [liquids, nasals, [s], [z]] [obstruents]

The proper treatment of nasal consonants became clear through testing of the initial version of the OT-based table. As noted above, significant weight is given to place features, but the productivity of place assimilation in nasals renders place less salient for transformations between nasals. Additionally, nasals are often confused for one another perceptually. Thus pairs of nasal phonemes are now scored as quite close to each other, despite significant differences in place. Scores between nasals and non-nasals were kept constant, though it was found that for scores between a sonorant and an obstruent, if the sonorant was a nasal, it was helpful to assign a slightly lower cost than if the sonorant was not a nasal.

The laryngeal phonemes of Arabic required some manual adjustment as well. Using true Arabic phonemes in the IPA variants produced by the Arabic rules was effective for the retrieval component. But not surprisingly, in ranking it was found that judgements of similarity by native English speakers tended to flatten out the distinctiveness of the Arabic phonemes. The nature of typical Roman transcriptions of Arabic names also contributes to this effect. Arabic laryngeals needed to be considered as more like their perceived English counterparts. Thus, for example, it was necessary that the velar stop [k] be set much closer to the uvular stop [q], flattening the salience of the place markedness chasm between these phonemes.

Finally, an overarching limitation is that, modulo the relative weighting implied by the general scalar framework outlined above, the exact value associated with each constraint violation is arbitrary. It would be preferable to substantiate the costs of constraint violation in a more systematic and theoretically or empirically justified manner.

6. Future Research

OT faithfulness constraints operating on ternary scales appear to provide additional formalization and justification for certain aspects of the cost model. However, the primary problem of explicitly quantifying constraint violation remains. The performance of the system will improve if additional research can produce a more grounded set of metrics. A number of lines of inquiry are available.

First, it should be possible to count marks, or violations, literally in a more convincing OT analysis of featural changes. The OT analysis could include universal faithfulness constraint hierarchies, operating in some cases on ternary scales. In addition to the IV scale, Gnanadesikan (1997) discusses possible ternary scales for vowel height and consonantal stricture. But the extension of the idea of ternary scales to other features, particularly place features, appears to be highly problematic. With place features, at least, we have recourse to the Universal Place Markedness hierarchy. A comprehensive OT analysis would no doubt be one that combines notions of faithfulness and markedness, a hallmark of work in OT. The focus must be on extending these ideas to all features.

Another source for quantification of constraint violation would be various probabilistic methods. In a corpus-based approach, large databases of names could be analyzed for the frequency of occurrence of specific phonological changes, such as de-voicing, and the analogous constraints could be quantified accordingly. This could be coupled with additional information, such as the statistics of phoneme confusion matrices available in the literature of telephony, or statistics describing the relative dominance of the dialects influencing the names of particular cultures.

Additionally, setting aside the issue of their merit as a theory of phonology, stochastic phonological grammars (Coleman and Pierrehumbert 1997) provide a potential source of useful data. In particular, this line of research centers on statistically valid measures of acceptability judgments of neologisms, which could be used to score the analogous violations of OT markedness constraints. Indeed, Coleman and Pierrehumbert argue against infinite constraint domination, with the claim that probabilities based on lexical distributions driving scalar judgments of acceptability are more psychologically plausible than single constraint violations that can cancel a parse. Judging the acceptability of a neologism seems qualitatively similar to judging the acceptability of, and relatedness among, personal names, as newly encountered personal names are effectively neologisms. The related work of Frisch (1997), which further incorporates data on speech errors and categorical perception, formalizes the notion of a gradient linguistic constraint, and in the process confirms the results attributing greater salience to initial sounds in judgements of similarity. In addition to being possible sources of useful data, the quantitative methodologies of these research areas may suggest ways to develop more rigorous methods for evaluating the ranking results of the system.

Finally, another possible source of data for quantifying constraint violation would be phonetics. Hayes (1996) argues for the induction of OT markedness constraints based initially on measures of articulatory difficulty. He argues that phonological grammars are phonetically grounded, but constraints are induced based on a synthesis of phonetic difficulty and a tendency toward formal symmetry. The idea is interesting, though only very limited data is presented.

7. Contact Information

Dr. John (Jack) Hermansen
Chief Technology Officer, IBM Global Name Recognition
(703)834-6200 x222 • jhermansen@us.ibm.com

Thomas Woodcheke
Worldwide Sales Manager, IBM Global Name Recognition
(703)834-6200 x253 • twoodche@us.ibm.com

Leonard Shaefer
Directory of Development, IBM Global Name Recognition
(703)834-6200 x228 • lshaefer@us.ibm.com

Timothy Paydos
Director of Marketing, Threat & Fraud Intelligence
(860)408-1639 • tpaydos@us.ibm.com

8. Additional Information

For the latest information about our products and services, see the following website: **www.ibm.com/**software/data/globalname/

9. References

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Coleman, John and Janet Pierrhumbert. 1997. *Stochastic Phonological Grammars and Acceptability*. Proceedings of SIGPHON 3.

Frisch, Stefan. 1997. *Similarity and Frequency in Phonology*. PhD Dissertation, Northwestern University. (Available at ROA).

Gnanadesikan, Amalia. 1997. *Phonology with Ternary Scales*. PhD Dissertation, University of Massachusetts, Amherst.

Hayes, Bruce. 1996. *Phonetically Driven Phonology: The Role of Optimality Theory and Inductive Grounding*. Proceedings of 1996 Milwaukee Conference on Formalism and Functionalism in Linguistics. (available at ROA).

Kenstowicz, Michael. 1994. *Phonology in Generative Grammar.* Blackwell.

Lombardi, Linda. 1998. Ms. (University of Maryland Working Papers in Linguistics).

Prince, Alan and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar.* Ms. MIT Press: to appear.

Stevens, S. S. 1946. *On the theory of scales of measurement*. Science 103(2684):677-680.

Lutz, Richard. 1997: *The Use of Phonological Information in Automatic Name Searching. In Proceedings of Symposium on Advanced Information Processing and Analysis*, Tysons Corner, VA, March 25-27, 1997.

IBM®

IBM (United States of America)
Entity Analytic Solutions
6600 Bermuda Rd, Suite A
Las Vegas, Nevada
United States of America, 89119

Printed in the United States of America
06-06

♻ Printed in the United States of America on recycled paper containing 10% recovered post-consumer fiber.

IBM's customers are responsible for ensuring their own compliance with relevant laws and regulations. It is a customer's sole responsibility to obtain advice of competent legal counsel as to the identification and interpretation of laws and regulations that may affect a customer's business and any actions required to comply with such laws. IBM does not provide legal, accounting or audit advice or represent or warrant that its services or products will ensure that a customer is in compliance with any law.

ON DEMAND BUSINESS™