PHONETIC TRANSCRIPTION STANDARDS FOR EUROPEAN NAMES (ONOMASTICA)

M. Schmidt, S.Fitt, C. Scott and M. Jack

Centre for Speech Technology Research,
University of Edinburgh, U.K.

ABSTRACT

This paper details the standards identified for phonetic transcription of names as part of the ONOMASTICA project, a European-wide research initiative for the construction of a multi-language pronunciation lexicon of proper names. The main design criteria adopted by the consortium for the development of this multi-language pronunciation dictionary are discussed, including aspects such as phonetic transcription standards, definitions of quality, quality control mechanisms and language specific details concerning phonetic transcription and the annotation of the language of origin.

Keywords: Multi-language dictionary of proper names; phonetic transcription standards; quality control.

1. THE ONOMASTICA PROJECT

The ONOMASTICA project was established as part of the 'European Commission Framework Programme - Linguistic Research and Engineering'. It seeks to create a set of pronunciation lexicons of European names, including city and town names, street names, family names, company and product names in a machine assisted fashion where expert phoneticians carry out editorial preparation of the project lexicon using customized software.

A total of nine languages of the European Community are covered in the project which include: Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish. The project thus has nine partners preparing the lexica for their respective languages from names data files provided by their associated telephone company.

The goal over the 2-year project is to derive pronunciation dictionaries for up to 1,000,000 names per language in a semi-automatic way and to investigate the problems of exchanging national names amongst the partners to create a matrix of 'nativised' pronunciations for each (thereby) foreign name in each other language.

1.1 OBJECTIVES AND EXPECTED IMPACT

The non-availability of large pronunciation dictionaries of names continues to impede the development of many applications in speech technology. In particular, the acceptability of applications where speech output systems provide spoken feedback depends heavily on the capability of producing correct, or at least acceptable, pronunciations for names of various categories.

The objective of the project is to make available, for widescale exploitation, quality controlled pronunciation lexicons in machine readable form (CD-ROM) for use in automatic language systems and of primary interest to international European companies in the telecommunications sector and in the European dictionary publishing industry. In particular, the multi-lingual dictionaries produced on this project will benefit products in the following sectors:

- Telecommunications: automated directory enquiry systems, reverse directory enquiry systems, catalogue ordering systems, telephone banking, automated credit card authorization, enhanced talking newspapers and books for the blind etc.

- Consumer sector: Map information and guidance systems, talking dictionaries and courseware systems for pronunciation teaching.

- Publishing: hard-copy as well as electronic dictionaries containing pronunciation fields.

2. SCIENTIFIC APPROACH

The conversion from an orthographic to a phonetic representation of a name in an automatic system can be achieved either by **dictionary** or by **rule**. The project aims to compile electronic dictionaries for names using machine assistance in the form of rule-based generation of name pronunciations as raw materials for expert phoneticians to edit as entries to the dictionary.

Although the main objective of the project is to provide a lexicon of names, one of its major goals is to develop an optimal set of grapheme-to-phoneme rules which function as an accelerator to human editing. The emphasis that is placed on human editing and automatic conversion by rule is expected to vary in different languages in the project because of obvious differences in the reliability of grapheme-to phoneme correspondences for different languages.

Historically, the development of rules for run time application has been preferred ([3];[5]), due to the capability of rules of treating unseen names, but the widespread availability of optical disk technology has greatly increased the feasibility of storing large dictionaries which could guarantee the correct automatic pronunciation of the vast majority of names in a national telephone directory if every person's name together with its phonetic representation were listed.

Furthermore, adopting the rule-based approach as the only method has its limitations due to the complexity of specifying grapheme-to-phoneme rules for names which can be very different from those of the general language. It is a well established fact [1],[6]) that grapheme-to-phoneme correspondences are different for names with different languages of origin, and it is also debatable whether the phonological systems of names are exactly equivalent to the phonological systems of those languages. The nature of the problem comes partly from the mobility of names, because names move with people and tend to surface in a language without passing through the slow linguistic process of borrowing and subsequent modification. Their anomalous pronunciations often fossilise and result in pronunciation difficulties for both man and machine.

3. CROSS-LANGUAGE PHONETIC CRITERIA

This section describes the standards agreed to by the consortium with respect to:

- Phonetic standards
- Quality specifications for lexicon entries

3.1 PHONETIC STANDARDS

3.1.1 Phonetic alphabets.

The final version of the lexicon will contain transcriptions coded as unique IPA numerical reference numbers as described in [2].

3.1.2 The level of transcription

The central purpose of the lexica is the provision of simple, comprehensible transcriptions which allow native as well as non-native speakers to produce adequate and natural pronunciations of names. Furthermore, the transcriptions should be usable (either directly or indirectly) as input for speech synthesis systems and/or as lexica in speech recognition applications.

Therefore, the level of transcription that has been agreed to be the most profitable for these purposes is a **broad phonetic** level. At this level very fine phonetic detail such as degrees of voicing in oral stops, degrees of aspiration in voiceless stops or assimilated vowel nasalization etc. are not transcribed. Important allophonic contrasts however, such as word final devoicing in German or clear and velarized /l/ in English as well as the contextually conditioned realization of the voiceless velar fricative in Greek and German ([x] before back vowels and [ç] before front vowels) are transcribed. For native speakers fine phonetic as well as allophonic contrasts are superfluous in a transcription due to their knowledge of the language. For non-native speakers fine phonetic detail adds unnecessary complications, whereas important allophonic contrasts are necessary in order to make adequate pronunciations.

3.1.3 Annotation of stress and syllabification

Lexical stress is marked on names which contain more than one syllable and phrasal stress is marked on compound names. Monosyllabic names are unmarked. Two levels of stress, primary and secondary, are marked by diacritic before the stressed syllable. Possible stress shift (in English) is also marked.

Syllabification and word boundaries are marked, following the principle of maximal syllable onset unless morphological considerations override this principle (see Section 4 for examples from English).

3.2 QUALITY SPECIFICATIONS

Each transcription in the ONOMASTICA database is assigned one of three quality bands, with Band I being the highest, enabling the user of the lexicon to determine the

reliability of the pronunciation (See section 3.2.1 below for pronunciation verification for English Band I.)

An initial goal of the project is to create a hand-transcribed set of 50,000 quality Band I names to be used as a basis for rules development and testing. To allow for maximum coverage this 'golden set' of high quality transcriptions will contain the most frequently occurring names. Transcriptions for the remaining names will subsequently be produced by rule, placing them in quality Band III. Quality Band III names will be checked and edited where necessary by hand and promoted to Bands I and II.

3.2.1 Verification of pronunciations (English)

It is an aim of the project to produce pronunciations which are not only acceptable to a native listener but also as far as possible to the owner of a name. For example, for English all pronunciations given a quality Band I are defined as being acceptable to the owners of the names. Many names in the set of English names provided by BT Laboratories will be familiar to the phonetician or can easily be checked in existing dictionaries, and pronunciations can immediately be verified or edited and assigned quality Band I. However, there is a significant number of names for which the pronunciation will not be known or for which there is an element of doubt or a possibility of alternative pronunciations. In order to provide acceptable pronunciations and so increase the number of quality Band I names in the database, various quality control procedures are being adopted.

One procedure is to contact the owners of names by telephone to confirm pronunciations. Contact telephone numbers are obtained from the BT 'Phone Disk' which operates on a PC. This enables the researcher to search for names with no need to specify an address or even a region. This is particularly useful for finding the owners of very unusual names which may occur only once and could be anywhere in the country.

Secondly, British schools and ethnic community groups are being invited to collaborate in the project. Teachers and community leaders are requested to provide information about unusual or commonly mispronounced names. Participants provide written annotations of such names by use of rhyming or reference to common words or parts of words to describe the pronunciation. Through this device large quantities of data containing unusual names, particularly foreign names, including information about their language of origin could become available

Finally, through the placement of advertisements and articles about ONOMASTICA in national newspapers, members of the public are being invited to submit details of their own unusual names.

4. WORKING PRACTICES

Working practices have been agreed for use of the project and are described here with specific reference to English.

4.1 MULTIPLE WORD ENTRIES

Multiple-word entries are included in the database and are transcribed in full, with the exception of recurring, predictable elements such as street name types (see section 4.2.2 below). This approach enables more accurate transcriptions to be given for certain names, such as 'Rowley', which is pronounced ['rou.li] in all cases except for the town 'Rowley Regis', which is [ˌrau.li 'ri:.dʒɪs].

4.2 STRESS

4.2.1 Phrasal stress

A single polysyllabic name can have both primary and secondary stress markers. However, in the case of multiple-word names a maximum of one stress per word is assigned, with only one primary stress which functions as a phrasal stress marker, for example 'Elim Pentecostal Church', which is transcribed as [ˌi:.lɪm pɛn.tɪˌkɒ.stɬ 'tʃɜ:tʃ].

4.2.2 Stress shift

For English, both primary and secondary stress are marked. Additionally, stress shift is marked on certain words, which enables more accurate prediction of stress in phrases. An example is 'Aberdeen', which in isolation is pronounced [ˌæ .bə'di:n], but is subject to stress shift when it precedes words taking primary phrasal stress, such as 'road'. In these contexts the main word stress shifts from the last to the first syllable, giving [ˌæ.bəˌdi:n 'roud] rather than [æ.bəˌdi:n ' roud]. 'Carlisle', on the other hand, is not subject to stress shift and would give [kɑ:ˌlaɪɬ ˌroud]. Since 'road', 'crescent' and so on are common elements in street names it is obviously more efficient to have a separate dictionary for these, to mark stress shift on individual lexical entries and to produce the combinations by rule, rather than having multiple-word entries.

4.3 SYLLABIFICATION

Many different methods of syllabification are possible and no one system is wholly satisfactory on all criteria -

phonological, morphological, acoustic, and articulatory. For syllabification of English transcriptions in this work, the principle of maximal onset is being used for simplicity, so that consonant clusters are treated as syllable initial if they are permissible clusters at word beginnings. 'Mostyn' is therefore transcribed as ['mɒ.stɪn] rather than ['mɒs.tɪn] or ['mɒst.ɪn]. However, this may be overridden by morphological considerations to give more intuitive syllabification, so that 'Foxcroft' is transcribed as ['fɒks,krɒft] rather than ['fɒk,skrɒft].

4.4 MULTIPLE PRONUNCIATIONS

Approximately 10% of the names transcribed so far have multiple pronunciations. All known possible pronunciations are entered within the criteria outlined above (for example differences due to surface phonetic realisations are not transcribed). The customized software used in the production of transcriptions enables the specification of information relating to category, language of origin and miscellaneous annotations. The following comments can be linked to specific pronunciations.

4.4.1 Category markers

In some cases two pronunciations differ in category, and so marking the category will aid the eventual user of the lexicon. For example, 'Clavering' as a surname is ['klæ.və.rɪŋ], whereas the town of the same name is ['kleɪ.və.rɪŋ].

4.4.2 Miscellaneous annotations

Sometimes pronunciations are annotated with respect to particular referents, for example the town 'Blean' in Kent is pronounced [blin], whereas 'Blean' in North Yorkshire is pronounced [bleɪn]. Another example is the surname 'Lamont', which has two different pronunciations, ['læ.mənt], which is the usual Scottish pronunciation, and [lə'mɒnt], which is used in Northern Ireland and is also used by the British politician Norman Lamont; this information is included as cross-indexed annotations in the lexicon.

4.4.3 Local variants

Pronunciations in accents of English other than RP are not transcribed, as this is outside the scope and aims of the project. However, where a local variant is markedly different in an unsystematic and unpredictable way, this is transcribed, for example [hʌn'stæn.tən], but there is also a local variant ['hʌn.stən] which is included in the lexicon and annotated as a local pronunciation.

5. CONCLUSIONS

The pursuit of onomastic research on a European scale permits novel cross language research concerning the pronunciation of names as well as the identification of languages of origin. The project is currently assembling a database of city and town names from non-border regions in each country, in order to train an n-gram based language identification system. This system allows the application of language dependent rule-sets for grapheme-to-phoneme conversion. The identification of non European languages is also part of this study, due to the large amount of non European names found in the telephone directories.

The project, in its later stages, will also see the exchange of the most common names in each language amongst all the partners, in order to construct a matrix of names pronunciations. This will be particularly interesting for the study of processes of nativization particularly with respect to the adaptation of 'foreign' graphemic or phonemic sequences to the language in question. This will be approached from two angles, firstly from the point of view of the native speaker of a language, and secondly, from the point of view of the adaptations that carriers of foreign names (or their descendants) make in order to assimilate the pronunciations of their names to a particular language.

REFERENCES

[1] Church, K. (1986). Stress assignment in letter to sound rules for speech synthesis. *Proc.ICASSP*, Vol. 4, pp. 2423-6.

[2] Esling, J. (1990). Computer coding of the IPA: supplementary report. *Journal of the International Phonetic Association*, Vol. 20, No. 1, pp. 22-26.

[3] O'Malley, M. and Caisse, M. (1987). How to evaluated Text-to-speech systems. *Speech Technology,* Vol. Mar./Apr., pp. 66-75.

[4] Pointon, G.E. (ed.) (1990). *BBC pronouncing dictionary of British names.* Oxford: Oxford University Press.

[5] Spiegel, M.F. (1985). Pronouncing surnames automatically. *AVIOS*, 10-12 Sep. 1985.

[6] Vitale, T. (1991). An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, Vol. 17, No. 3, pp. 257-76.

[7] Wells, J.C. (1990). *Longman pronunciation dictionary.* Harlow: Longman.