# THE ONOMASTICA INTERLANGUAGE PRONUNCIATION LEXICON

*The Onomastica Consortium**

## ABSTRACT

This paper presents one of the linguistic resources developed in the scope of the ONOMASTICA European project. In terms of size, the interlanguage pronunciation lexicon represents a very small fraction of the global lexicon produced by the project. The interest of this research tool, however, derives from its particular contents: 1,000 names from each of the 11 languages represented in the consortium, with corresponding phonetic transcriptions in all the languages. The exchange of names amongst the partners was done for the purpose of making "nativised" pronunciations for each name. The paper describes in detail the contents of the matrix lexicon and discusses the problems encountered in defining these nativised pronunciations.

## 1. INTRODUCTION

The recently finished ONOMASTICA project was a European wide research initiative within the scope of the Linguistic Research and Engineering Programme, whose aim was the construction of a multilanguage pronunciation lexicon of proper names [6]. The project covered eleven European languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. Eleven associated partners from telephone companies have provided data files including names of persons, cities, towns, streets and companies. One of the main goals of the project was to derive pronunciation dictionaries for up to one million names per language in a semi-automatic way.

In general, the performance of grapheme-to-phoneme conversion systems for proper names is much worse than the one observed for the common lexicon. This fact is not surprising since in most languages the names may obey to different morphological and phonological rules compared to ordinary words. Part of the problem derives from the mobility of names, as they move with people from one country to another, showing different degrees of adjustment to the sound structure of the language in which they surface. Other sources of difficulty can, however, be found. The orthography of last names can be rather conservative and, as it does not conform anymore to the general orthographic rules, its phonetic interpretation is sometimes misleading. Furthermore, some applications imply the ability of generating correct pronunciations for acronyms which, for some languages, can follow rules significantly different from the ones observed for the common lexicon.

Part of the work in this project was therefore aimed at upgrading existing rule engines to cope with the problems posed by proper names. A significant part of the work was also devoted to the development of self-learning grapheme-to-phoneme conversion methods and the comparison of their performance with the one of rule-based methods. These self-learning approaches included both conventional backpropagation and self-organizing neural networks, as well as various symbolic learning techniques, ranking from analogy-based learning to table look-up. This latter approach, developed by CPK [1] was tested by most of the partners, allowing inter-language assessment of its performances.

The number of entries in the ONOMASTICA lexicon significantly differs from language to language, ranging from one hundred thousand to more than one million. These were all automatically processed to provide broad phonetic transcriptions. A large percentage of these transcriptions was manually verified by at least one trained phonetician, who provided up to 5 alternative pronunciations for each entry, tagged with the corresponding category (first name, surname, company name, street name, town name, and region name) and in some languages where this information was available with its etymology and frequency of occurrence. Quality assurance measures have played a key role throughout the project. Thus, three quality bands have been identified, depending on the certainty of the transcriptions (I: verified by a transcriber who is certain of its correctness; II: verified by a transcriber with some uncertainty; III: not verified). One thousand entries from each band have been randomly selected and their correctness was judged by independent auditors from each lan-

*The work on the interlanguage lexicon involved the following researchers (in alphabetical order): Ove Anderson (CPK), Lou Boves (Univ. Nijmegen), Paul Dalsgaard (CPK), Vassilis Darsinos (Univ. Patras), Bjorn Granstrom (KTH), Joakim Gustafson (KTH), Henk van den Heuvel (Univ. Nijmegen), Mervyn Jack (CCIR), George Kokkinakis (Univ. Patras), Emmy Konst (Univ. Nijmegen), Michael Logothetis (Univ. Patras), Isabel Mascarenhas (Center of Linguistics of Univ. Lisbon), Andreas Mengel (Institut für Fernmeldetechnik), Peter Molbaek (CPK), Georg Ottensen (Sintef Delab), Jose Pardo (UPM), Vito Pirrelli (ICL), Mark Schmidt (CCIR), Andrew Sutherland (CCIR), Isabel Trancoso (INESC/IST), Francisco Valverde (UPM), Céu Viana (Center of Linguistics of Univ. Lisbon), and François Yvon (Telecom Paris). *The contributions to this paper have been collected by Isabel Trancoso, INESC (imt@inesc.pt).*

guage. Part of this ONOMASTICA pronunciation lexicon, which totals 8.5 million European names, is included in a CD-ROM with currently 25,000 band I entries from eight languages.

Another important goal of this project was to investigate the problems of exchanging national names amongst the partners to create a matrix lexicon of 'nativised' pronunciations for each foreign name in each other language. This inter-language lexicon, which is also included in the above mentioned CD-ROM, is the subject of the present paper. We shall start by briefly describing its design criteria and contents. The second and largest section of the paper is devoted to discussing "nativised" pronunciations and the various factors that can influence them. Before concluding, we shall briefly illustrate the use of the application programmer's interface (API) to the ONOMASTICA inter-language database.

## 2. LEXICON DESIGN CRITERIA AND CONTENTS

Whereas the set of the 11 national pronunciation lexicons is directly suited to immediate exploitation, particularly in the development of telecommunications applications, the inter-language lexicon should be viewed rather as a research tool. It is limited to 1000 names per language and therefore contains 11,000 entries, with eleven transcriptions each, amounting to a total of 121,000 transcriptions.

The design criterion for this database was primarily to emphasize the potential of use of this type of lexicons in multi-lingual speech recognition applications involving users in different European countries. With this in mind, the selected vocabulary was restricted to names of cities, towns, airports, stations, monuments and landmarks whose size, historical significance, and geographical importance (in terms of transport, namely) justify their inclusion in touristic guidebooks. The targeted applications are the ones which are most likely to be used by non-native users, thus implying the recognition of considerably different pronunciations: travel information, flight booking, weather forecasting, road report systems, etc.

Table 1 specifies the contents of the matrix lexicon exchanged among the partners in terms of categories. The first category includes names of cities, towns, regions and islands. For some languages, this is the only category present. For others, several criteria have been used to restrict the number of towns (size, administrative role, touristic importance, or even association with famous products like cheese and wine). The second category includes names of rivers, lakes, bays, channels, mountains, volcanoes, capes, gulfs, caves, and other geographical landmarks. There is naturally some overlap between the two categories, as, for instance, the most important rivers may be also region names. The third category includes churches, museums, bridges, towers, palaces, spas and other sites of touristic interest. For the most important cities of each country, names attached to a part of the city are also included. These are frequently names of train and metro stations, streets, squares, parks, etc. which may be associated with the closest monument. For some romance languages, a large percentage of this category are religious names. Famous paintings and other national art treasures may also be included here. The final category includes miscellaneous information: touristic events, gastronomy, names of foreign cities, etc.

| CATEGORY | I | II | III | IV |
|---|---|---|---|---|
| de | 100 | 0 | 0 | 0 |
| dk | √ | √ | √ | √ |
| es | 26 | 43 | 22 | 8 |
| fr | 91 | <1 | 8 | 2 |
| gr | 100 | 0 | 0 | 0 |
| it | √ | √ | √ | √ |
| nl | 74 | 4 | 22 | 0 |
| nw | √ | √ | √ | √ |
| pt | 58 | 2 | 40 | 0 |
| se | √ | √ | √ | √ |
| uk | √ | √ | √ | √ |

**Table 1:** Percentage of entries in each category for the 11 languages. √ - percentages not available

## 3. DEFINING NATIVISED PRONUNCIATIONS

One of the most interesting aspects of the work on the inter-language matrix lexicon consisted in defining "nativised" pronunciations. Many different criteria could be adopted in this definition. The default nativised pronunciation selected by the consortium is the one of a native speaker with little past exposure to foreign languages. Generally, this default nativised pronunciation in each language closely follows the transcription generated by the grapheme-to-phoneme rules for that language. The rule set, however, must be modified in many cases to take into account characters with diacritics which are not present in the language and unfamiliar grapheme sequences.

### 3.1. Sources of variability

Between the default transcription and the "native" one provided by the original partner, a wide range of possible pronunciations can be found. Additional transcriptions can thus be optionally provided to reflect increasing degrees of exposure to foreign languages. This wide range of transcriptions is due to the interplay of several factors. One of the factors is the reader's ability to identify the name as foreign by its orthography. In fact, many foreign names may not be identified as such because their orthography conforms to the phonotactic constraints of the native language. On the other hand, some non-existent grapheme sequences may lead to a name being identified as foreign, but not to the correct identification of its origin. This is one aspect where the two parts of

the ONOMASTICA lexicon, the interlanguage and the national lexicons, may differ. In fact, the etymology of the name in the national lexicons is not generally known, and the task of guessing is left to the transcriber.

Even if one assumes a correct recognition of the origin of a name, there are many other factors such as the knowledge of the pronunciation rules in the foreign language, the knowledge of the actual local pronunciation of the proper name and the ability to pronounce the sounds of the foreign language. The first two factors are related to what we will designate in this context as the reading competence of the subject and the third with his/her pronunciation competence [5]. Reading competence is dependent on the affinity between the target language and the native one (for instance, whether they belong to the same Germanic or Romance language group), and also on the familiarity of the native speaker with the target language (English and French, for instance, are taught in secondary school in many European countries). When the speaker completely ignores the pronunciation rules of a foreign language, he may typically look for similarity features in the languages known to him/her in order to choose the pronunciation.

### 3.2. Choosing sound inventories

The pronunciation competence concerns the ability of a speaker to pronounce sounds which do not exist in his/her native language. Many of these sounds are approximated with native ones. For instance, the French nasal vowels are approximated in Norwegian (where they do not exist) with the vowel followed by a nasal consonant; the [œ] and [ø] sounds are commonly replaced by [ɐ] in Portuguese [7], etc.

In some cases, however, the phone set of the primary language is expanded to encompass some phones from other languages. In Italian, for instance, 5 new symbols were added to the original phonetic alphabet: [ʒ] (to transcribe the *j* in French, as in *journal*); [h] (for the Spanish *j*, as in *Julio*; [y] (for the French *u*, as in *Durand*); [œ] (for the first vowel of *Voeller* in German); and the schwa [ə] for dealing with two separate phenomena: analogising some foreign sounds such as in the pronunciation of *de* in French, and as a dummy vowel to be inserted when needed to pronounce otherwise non pronounceable syllables, as in the pronunciation of French *Argenteuil*, where a schwa is added to word final [j] for a pronounceable syllable to be created.

It is interesting to notice that the letter *j* in initial position gets quite distinct transcriptions in the different languages. This was observed in a study of 5 languages reported in [3] (Swedish, English, French, German and Italian). Hence the addition of new phonetic symbols to transcribe this letter in foreign languages was adopted by several other partners as well.

### 3.3. Choosing the context

Another important aspect is context: the pronunciation may be more or less nativised depending on the person one is talking to and the situation. For instance, when talking to a person with little knowledge of the foreign language one is using, the pronunciation tends to be strongly nativised, even in cases of good pronunciation competence. A good example of the type of context one could imagine for producing the default nativised pronunciation is the following: "You read about a place in a traveling guide, and this place is unknown to you. The country is known, but you cannot speak the language. You call your local travel agency and say: *I would like to go to ...* What would be your pronunciation ?".

The combination of different degrees of reading and pronunciation competence causes a wide range of possible pronunciations, as mentioned above. Theoretically, however, it is interesting to define a hypothetical native speaker who has full knowledge of the spelling conventions of foreign languages, but is restricted to the phoneme set of his/her native language. This second optional "nativised" pronunciation was provided by some partners for some of the languages. The comparison between the different nativised pronunciations of a single name in the different languages is currently in progress and we expect it to provide interesting clues about the affinities between the target and primary languages. Also interesting is the comparison between the default nativised pronunciation assuming null reading and pronunciation competence, and this second nativised pronunciation assuming full reading competence.

This comparison was done for a subset of names (250) from five languages selected to reflect different degrees of familiarity and affinity with the primary language (Dutch, in this particular study [4]). The familiar languages were German, French and English, which are taught in school, and the unfamiliar ones are Swedish and Italian. The default nativised pronunciations were generated by rule and then compared to the ideal "Dutchised" ones. In order to align the transcriptions, a dynamic programming algorithm was adopted to find the optimal match between transcription strings. The algorithm produced minimised cumulative distance scores which were later submitted to an analysis of variance. English and French achieved distance scores greater than 1 (1.4 and 1.7, respectively), which means that the Dutch grapheme-to-phoneme rules generate output which is far from the ideal Dutchised pronunciation for these two languages. For Swedish and Italian, the matching is better (0.6 and 0.7, respectively), and best results were achieved for German (0.4). This tends to show that the affinity between the target and primary languages seems to play a major role.

### 3.4. Narrowing the phonetic transcription

The placement of syllabification marks in the transcriptions provided in the inter-language matrix lexicon also posed some interesting problems, although it was not considered mandatory for this corpus. In fact, different syllabification criteria were adopted to process the eleven languages, raising problems for instance when a name from a foreign language for which syllabification criteria are strongly dictated by morphological structure must be nativised in a language which has different syllabification criteria. Syllabification errors due to lack of knowledge of the foreign language morphology may occur, independently of the criteria used for the native language.

The ONOMASTICA project aimed at broad phonetic transcriptions, not necessarily including prosodic structure. Narrower phonetic transcriptions including for instance lenition phenomena could be optionally provided. Notice also that a large percentage of the orthographic entries of the interlanguage matrix lexicon include several words per entry (e.g., *Aix en Provence*, which implied the application of inter word coarticulation rules).

## 4. THE INTERLANGUAGE CD-ROM

The interlanguage matrix lexicon forms the last directory included in the ONOMASTICA CD-ROM, together with the 11 national directories. Although each of the partners used its own machine-readable phonetic alphabet for transcribing both the national and interlanguage entries, the phonetic transcriptions included in the CD-ROM have been translated into the International Phonetics Association Standard Computer Coding [2].

An application programmers' interface has been developed to provide a convenient method to access the data held on CD-ROM. Written in C, it can be used either from DOS or Windows, offering the basic functions to *open*, *search*, *read*, and *close* a data file. A Visual Basic program has also been developed to demonstrate the use of the API calls.

## 5. CONCLUSION

The paper presented the inter-language pronunciation lexicon produced by the ONOMASTICA project. The matrix lexicon includes 1000 touristically important names from each of the 11 countries of the project, with crossed nativised pronunciations in each of the languages. We have emphasized the factors influencing nativisation, and compared different degrees of adjustment to the sound structure of foreign languages. Although the current project ended in June, the work on ONOMASTICA will continue at least until 1997 with the introduction of new partners, addressing the names of Eastern and Central European names - Czech, Estonian, Latvian, Polish, Romanian, Slovakian, Slovenian and Ukrainian, in a new project funded by the EC Copernicus Programme.

## REFERENCES


[1] O. Andersen and P. Dalsgaard, "A Self-Learning Approach to Transcription of Danish Proper Names", Proc. ICSLP'94, Yokohama, Sept. 94, pp. 1627-1630.

[2] J. Esling, "Computer coding of the IPA: Supplementary Report", Journal of the International Phonetic Association, 20:1, 1990, pp. 22-26.

[3] J. Gustafson, "Transcribing names with foreign origin in the Onomastica Project", Proc. Int. Congress on Phonetics, Stockholm, 1995.

[4] H. van den Heuvel, "Pronunciation of foreign names by Dutch grapheme-to-phoneme conversion rules", Proc. of the 2nd Onomastica Research Colloquium, London, 1994, pp. 87-93.

[5] A. Mengel, "Transcribing names - a multiple choice task: mistakes, pitfals and escape routes", Proc. of the 1st Onomastica Research Colloquium, London, 1993, pp. 5-9.

[6] M. Schmidt, S. Fitt, C. Scott and M. Jack, "Phonetic transcription standards for European names (ONOMASTICA)", Proc. of the European Conf. on Speech Technology, Berlin, 1993.

[7] M. C. Viana, I. Trancoso and F. Silva, "On the Pronunciation of proper names and acronyms in European Portuguese", Proc. of the 2nd Onomastica Research Colloquium, London, 1994.