Issues in the pronunciation of proper names: the experience of the Onomastica project

Isabel Trancoso
INESC / IST
INESC, R. Alves Redol 9, Lisbon, Portugal
E-mail: Isabel.Trancoso@inesc.pt

With the cooperation of *Céu Viana*(CLUL, Lisbon)

Abstract

This paper discusses several relevant issues concerning the pronunciation of proper names. Although it was motivated by the experience of the ONOMASTICA European project, it reflects a personal view, not constituting, therefore, an official document of the project.

The major outcome of the project was the production of two important linguistic resources: the set of 11 national pronunciation lexica and the inter-language pronunciation lexica. The paper starts by a description of the goals of the consortium, followed by the format and contents of the two lexica. It then addresses the problem of automatic grapheme-to-sound conversion, whose application is almost mandatory for the development of such large-scale resources, and describes and compares several of the methods adopted during the project. The two types of lexica are discussed, with an emphasis respectively on the pronunciation of acronyms and the nativization of the pronunciation of foreign names.

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 multi-language pronunciation lexicon of proper names [8]. The project covered eleven European languages, with corresponding academic and associated partners in each language: Danish (CPK, Univ. Aalborg + Jydsk Telefon), Dutch (Dept. Language and Speech, Nijmegen + PTT Research), English (CCIR, Univ. Edinburgh + BT Laboratories), French (ENST, Paris + France Telecom), German (Inst. für Fernmeldetechnik, Berlim + Deutche B. Telekom), Greek (Dept. Electrotechnical Engineering, Univ. Patras + Intrasoft), Italian (Inst. de Ling. Computacional, Pisa + CSELT), Norwegian (SINTEF DELAB, Trondheim + Norwegian Telecom Research), Portuguese (INESC, Lisbon (with CLUL) + Portugal Telecom (formerly TLP)),

Spanish (UPM, Madrid + Telefónica) and Swedish (Kungl Tekniska Hogsk., Stockholm + Telia (Infovox)). The associated partners, which were mostly from telephone companies have provided data files including names of persons, cities, towns, streets and companies. The prime contractor was CCIR.

In general, the performance of grapheme-to-phone 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.

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. Part of the work in this project was therefore aimed at upgrading existing rule engines to cope with the problems posed by proper names. In this scope, a significant effort was devoted to the development of self-learning grapheme-to-phone conversion methods and the comparison of their performance with the one of rule-based methods. For some of the languages, the pronunciation of acronyms and the way it deviates from the pronunciation of the common lexicon was also a relevant research topic. 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 language.

This paper addresses all the above mentioned issues, starting with a brief discussion of the contents and format of the two major linguistic resources produced by the consortium: the 11 national pronunciation lexica and the inter-language pronunciation lexicon. It then discusses briefly some self-learning approaches for automatic grapheme-to-phone conversion, the problems raised by the pronunciation of acronyms, and the factors influencing nativised pronunciations. The paper will be somewhat biased in the sense that many of the examples used throughout the paper concern the Portuguese language.

2 THE ONOMASTICA PRONUNCIATION LEXICA

2.1 The 11 national lexica

The number of entries in the ONOMASTICA lexicon significantly differs from language to language, ranging from one hundred thousand to more than one million. One obvious cause for these differences is the size of the population in each country. However, it is also important to notice that several countries processed isolated as well as compound entries (i.e., St. Paul's Cathedral was considered a single entry), whereas other countries only processed entries formed by single words.

All the entries were 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: quality I (verified by a transcriber who is certain of its correctness), quality II (verified by a transcriber with some uncertainty), and quality III (not verified). One thousand entries from each band have been randomly selected and their correctness was judged by independent auditors from each language.

For the languages in which the information on the frequency of occurrence of each entry was available, a study of the distribution of names and their corresponding coverage was made, providing interesting results. For instance, the Portuguese lexicon corresponding to the two largest cities includes 100,000 different entries, of which about 50% are unique occurrences (corresponding mostly to very small companies, foreign names and many spelling errors). Only less than 3% of the entries occurs more than 100 times and roughly 13% occur more than 10 times. With only this last subset of entries, one could achieve a coverage of about 84% of the full names in the national directory. The representativity of the ONOMASTICA lexicon, therefore, is very high.

2.2 The interlanguage pronunciation lexicon

Whereas the set of the 11 national pronunciation lexica 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 11 transcriptions each.

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 targeted at touristic applications, and included four main categories of names:

- cities, towns, regions and islands;
- rivers, lakes, bays, channels, mountains, volcanoes, capes, gulfs, caves, and other geographical landmarks.
- 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.
- miscellaneous information: touristic events, gastronomy, names of foreign cities, etc.

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.

2.3 The ONOMASTICA CD-ROM

Part of the ONOMASTICA pronunciation lexicon, which totals 8.5 million European names, is included in a first CD-ROM with currently 25,000 band I entries from eight languages, together with the interlanguage lexicon. For the sake of uniformity amongst languages, no frequency of occurrence, neither etymology information was included. 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]. The 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 compound entries (also very frequent in the interlanguage lexicon) imply the application of inter word coarticulation rules (e.g., *Aix en Provence*).

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.

3 SELF-LEARNING GRAPHEME-TO-PHONE CONVERSION

Although most of the partners had previously developed sets of rules for automatic grapheme-to-phone conversion which had to be upgraded for dealing with proper names, the largest effort invested in this area was in the development and testing of self-learning methods. These 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.

The work on upgrading rule sets for dealing with proper names varied significantly from language to language. For Portuguese, for instance, comparative tests of the rule-based method with a subset of the common lexicon, containing about 8000 words and with a corpus of the most frequent 15,000 proper names (excluding acronyms), yielded a relatively small difference (5% vs. 7% word transcription errors, respectively). This proves that in Portuguese, contrarily to what is referred for some other languages, the letter-to-phone correspondence does not significantly differ for the two corpora. Notice, however, that the test corpus of proper names includes only the most frequent names and that slight modifications of the rules are needed to take into account some characteristics that cannot be found in the two test corpora: the occurrence of some geminate consonants (*tt*, *ll*, *mm*, etc.) in old spellings of names and the occurrence of uncommon final consonants in foreign names and acronyms. We shall deal with these particular problems in later sections.

3.1 Neural Networks

The application of neural networks to letter-to-phone conversion dates back from 1987, when Sejnowski and Rosenberg presented the NETTALK system [9]. As in this pioneering work, the type of network adopted by the Portuguese team is a conventional multi-layered, feedforward neural network, trained by the backpropagation algorithm [10]. The learning phase is preceded by an automatic alignment procedure, which yields about 200 different letter-phone combinations. Phonemic nulls are inserted to account for graphemes with no phonemic realization (the initial "h" in Portuguese, for instance). The definition of graphemic nulls is also possible, although we avoided it by the alternative definition of new symbols for compound phones.

Each network input pattern is based on one grapheme and its context provided by nearby graphemes. The desired network output is the phone aligned with the input character. Several network architectures and context lengths have been tested. The one which yielded best results so far is illustrated in Fig.1. The input layer consists of 11 clusters of neurons, one cluster for each grapheme: the one to be transcribed, 3 graphemes to its left and 7 graphemes to its right, from which only 5 are used for phonetic transcription, the two last ones being exclusively used for stress assignment. Each grapheme is encoded by a group of 36 neurons. The hidden layer is

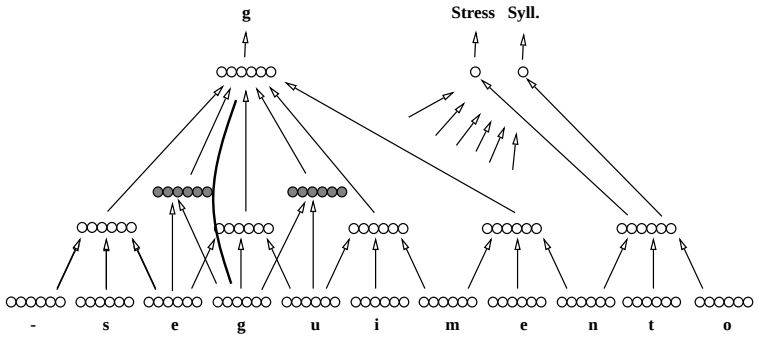

Figure 1: Architecture of the multi-layered neural network

structured into 5 clusters of trigraphs and two clusters of digraphs, each consisting of 20 neurons. There are 47 output neurons, one for each of the 45 different phones (including the phonemic null and the compound phones), and two to encode the primary stress mark and the syllable boundary. Shared weights have been adopted in order to reduce the number of weights to be adjusted. There is also one direct connection betwen input and output.

The network has been trained with a subset of the common lexicon not used for testing, which contains approximately 100,000 phones. The error percentage at segment level reached 1% at the end of 40 iterations. Tests with the two above mentioned corpora yielded an error percentage which is only slightly lower for the common lexicon than for proper names (7% versus 12%). It is interesting to notice that a significant percentage of the words in which the rule-based method fails was also wrongly transcribed by the neural network approach (74% of the cases for the proper names corpus, and 59% for the common lexicon). Moreover, roughly half of these wrongly transcribed words are transcribed in the same way buy the two approaches. As in the rule-based approach, most of the transcription errors concern the graphemes *e*, *o* and *x*. The network, however, shows some additional difficulties in dealing with vowel nasalisation, vowel raising and diphthongisation. The nasalisation difficulties are avoided by the rule-based method by placing syllable marks prior to phonetic transcription, and the vowel raising ones are avoided by also performing stress assignment before the transcription phase. Some of the generalisation problems are due to the small representativity of the corresponding grapheme sequences in the training corpus. Many interesting aspects of the performance of this neural network remained to be explored, namely the analysis of the activation patterns to determine functional groupings.

3.2 Table Look-up

The table look-up approach was developed by the Danish academic partner in ONOMASTICA (CPK) who supplied the software package SELEGRAPH which was tested for the different languages [1]. The main difference between this type of self-learning approach and the neural network approach described before is the complete lack of generisation capabilities in table look-

up approaches, a disadvantage that is to some extent counter-weighted by their much faster training procedure. Table look-up approaches are trained on the basis of paired grapheme-phoneme strings, dynamically determining which left and right contexts are minimally sufficient to be able to map any of the graphemes to the correct phone with absolute certainty.

The table look-up training is preceded by two phases: the alignment phase, just as for the neural network approach, and the computation of mutual information, to determine how many context graphemes to include and the ordering in which the context graphemes should be considered. The training results in tree-structured statistics for each grapheme in a given context, of the number of occurrences of each possible phoneme. Default mappings are used for ambiguous grapheme-to-phoneme conversions and unseen words with grapheme sequences not present in the training corpus.

As for the neural netwok method, the table loop-up approach was trained for Portuguese with a subset of the common lexicon, and tested with both common lexica and proper names. An analysis of the errors showed the same type of difficulties as faced by the neural network, although much more frequently, which evidences its lack of generalization capabilities.

These two self-learning approaches showed the potential to perform as well as the rule-based ones for our language provided larger training corpora are used, and/or separate processing of syllabification and stress assignment is adopted.

3.3 Analogy-based

Analogy-based approaches were developed namely by the French and Italian academic teams. This type of approaches also assumes an aligned training corpus of orthographies and transcriptions. For each test word, the pronunciation is analogised to known pronunciations through the application of two functions: a mapping function, defined over character strings, by which the orthography of the test word is projected onto training words in order to select the best candidate analogue(s) among them, and a recombination function, defined over phonological strings, which pastes together the transcription of the substrings that the selected analogues share with the test word [6]. It is well known that children learn how to read written words aloud by some form of analogy-based reasoning, but there is still an ample area for research on what is the role and function of analogising factors, and on how to find flexible and yet computationally tractable mapping and recombination functions. Despite these difficulties, the results obtained for Italian and English proper names have surpassed both rule-based and table look-up approaches, although the task is more challenging for the latter language.

4 ACRONYMS

For some of the languages in the project, the data provided by the industrial partners contained company names as well as names of persons, streets and towns. The pronunciation of acronyms, which constitute a very significant part of the set of company names, was one of the research topics followed in particular by the French and Portuguese teams [10].

In the Portuguese lexicon, acronyms constitute 38% of the most frequent 50,000 entries of the original database. The performance of both the rule-based and the self-learning methods for this category drops significantly (only 57% and 49% of the corresponding phonetic transcription coincide with the manually corrected ones). Moreover, their pronunciation by native speakers also shows considerable variation. These two facts motivated a detailed study of the lexical formation processes of acronyms and their relationship with the variability of pronunciation, which will be the subject of the next two subsections. One class of acronyms (*siglae*) deserves special emphasis, as they may be either read or spelled, and will therefore be dealt with in the last subsection.

4.1 Lexical processes

In the building up of company names, common affixes, words, roots, first names, last names, toponyms and almost any possible truncation of those are used, combined with each other, with foreign words, and with word endings specific to this category. For Portuguese, for instance, we have identified about 700 such constituent elements which covered roughly 65% of the acronyms. Several types of typical lexical formation processes can be found in our lexicon: acronymy (in the strict sense), blending, *siglae* (or abbreviation proper) and, although rarely, single truncation. In most cases, they involve abbreviations of the general designation of the company or of one or more names/surnames of its owner(s). These abbreviations may include only the initial letter of each one (*siglae*), one or more letters, syllables or even initial morphemes (acronyms in the strict sense) or any sequence of selected elements (blending). The fundamental distinction is not in the number of letters retained, but rather in the criteria which is on the basis of their selection: whereas acronyms are created to be "read", *siglae* may be either read or spelled, frequently being adopted just for the sake of easy writing.

However, a large percentage of acronyms in the Portuguese lexicon is formed by other types of lexical processes, such as compounding, which are not so frequent in the common lexicon of our language. Portuguese orthography treats compounds in a way that may be ambiguous from a morpho-syntactic or semantic point of view, but whose main goal is ensure a correct reading. From this point of view, the fundamental distinction is between word and root compounds [11]. The first may have as many non-raised vowels as their constituents whereas the latter also

have, besides those, a binding vowel, /i/ or /ɔ/ which, in the second case, is not raised either. Graphically, these two types are distinguishable by the fact that the first ones are typically written as separate words (often with hyphens or spaces), and the second ones are written as a single word. Therefore, word-compounds do not need any special treatment by the rule system and most root-compounds can be easily identified on the basis of a reduced list of bound morphemes, typically of Greek or Latin origin. For acronyms, however, this type of processing is clearly inadequate since, regardless of its type, each compound is graphically coined as a single word and the stress marks of the first element are never retained.

4.2 The pronunciation of acronyms

In order to study the variability in pronunciation among speakers and relate it with the lexical processes used in the formation of the company name, two types of complementary information sources were used: (1) direct contact with a set of companies in order to find the origin and intended pronunciation of their names; (2) a reading test using 10 subjects of university background who were asked to read a list of 100 randomly selected acronyms, almost totally unknown from the speakers. The direct contact with companies mainly showed the large variability of criteria that can govern the choice of a name: one can choose if the resulting form should sound native or foreign, if it should be homograph or homophone of a common lexicon word or totally distinct from these; one can also favour and penalize certain semantic associations or simply avoid that the chosen name be identical or very similar to an existing one. The intended pronunciation, however, is not always the one which is most frequently adopted by native Portuguese speakers. Very often, multiple acceptable pronunciations are found, and it is hard to distinguish the most probable one, but at least one should detect which pronunciations are clearly unacceptable. The fact that, in the above mentioned reading test, only 37% of the forms were pronounced identically by all speakers clearly demonstrates the extreme variability these forms are subject to. A careful analysis, however, shows that this variation is not random.

Many of the forms in our corpus are always analysed as compounds, as *globomar* and *frangolandia*, for instance. Since the binding vowel "o" in root-compounds is identical to the male gender mark of word-compounds, this class of forms is inherently ambiguous. *Globomar*, for instance, was pronounced as [glo.bɔˈmar] by 40% of the speakers and as [glo.buˈmar] by 60%, but other realizations are clearly unacceptable. These oscillations suggest that the recognition of words or roots within words is not part of the task of reading Portuguese. If it were, then forms such as *alfasom* would be invariably treated as word compounds (*alfa + som*) and pronounced as [aɬ.fɐˈsõ], as originally intended. However, in 60% of the cases, this form is treated as a single word and the inter-vocalic "s" pronounced as [z], as dictated by the general rules. The diffi-

culty of recognizing words or roots within words suggests the accuracy of the type of processing adopted for our rule system, which includes the search for an initial element, in general bi or trissylabic, that ends in /i/ or /ɔ/ and may be interpreted as a root compound. The previous example is also one of many acronyms that do not conform to the basic orthographic rules in what concerns the grapheme-to-phone correspondence. The omission of diacritics indicating the stress position is also very frequent in acronyms.

Most people are aware that company names differ from the common lexicon and proper names, in orthography as well as in pronunciation. Thus, as they become more conscious of the class of names at stake, they tend to analyse every form as compound, assigning stress to each element that may coincide with a root or word or which may be interpreted as a truncation of either of these. Since stressed vowels do not undergo reduction or deletion, sequences of syllables with open vowels become very frequent. This may explain the appearance of a general strategy of not allowing vowel raising in pre-stressed position. This strategy is systematically adopted in all cases which have endings that characterize this class of names (e.g., *-ax, -ux, -trans, -tur*). Unstressed vowels in word final position, however, always undergo reduction or even deletion.

4.3 Reading and spelling of *siglae*

Siglae pose particular pronunciation problems. Some *siglae* are mandatorily read, some are spelled and some may be pronounced in both ways. Although not so frequently, some may even have mixed pronunciations - partly read and partly spelled. To choose between these alternative options is one of the main problems with this class of names.

In its previous version, our rule system naturally spelled all *siglae* with no vowels and tried to read all that had at least one vowel. The last condition is obviously a necessary condition for reading, but it is not sufficient: in our lexicon, about half of the *siglae* which are spelled contain at least one vowel.

The length of the sequence is one of the important features: with rare exceptions, *siglae* with less than 3 characters are spelled and the ones with more than 5 are preferably read. The two basic modes are possible with sequences of intermediate length (3 to 4 letters), but one cannot typically be used in place of the other. Certain patterns, such as CVCV are always read (e.g. *FIFA*) and others such as VCCC are spelled (e.g. *APDC*). With very rare exceptions, the CVC *siglae* are read (e.g. *CAP*); however, not all *siglae* that contain two vowels, such as VCV or CVV are read (e.g. *IPE*).

This type of observations has been made for other languages, and has been on the basis of some attempts to explain the pronunciation modes adopted for *siglae* as a function of the interaction of different prosodic features. Hence, for instance, in order to be read, each segmental

sequence must allow a syllabic analysis in agreement with the set of general principles and with the specific restrictions of the language. However, it must also correspond to a possible word pattern, both in extension and weight. In some cases, however, the structure and weight restrictions may not be compatible and the resolution of this conflict depends of their relative importance. Plénat [7] proposes minimum and maximum weight thresholds for a *sigla* to be read in French and refers some examples of possible conflicts. The minimum threshold of two morae (corresponding to a monossylable with branching rhyme or to a dissylable), defines a limit below which *siglae* are mandatorily spelled, and the maximum threshold of three syllables defines another limit above which they are mandatorily read. These syllabic weight restrictions coexist with a set of structural restrictions which determine the spelling of *siglae* whose constituents may be considered ill-formed as, for instance, a foot with a hiatus or without any CV syllable. The hiatus may be the cause for spelling *siglae* with a CVV pattern, since the resulting form does not overcome the three-syllable threshold. When the number of syllables overcomes this maximum threshold, the hiatus is tolerated and *siglae* are preferentially read. The *siglae* which allow both pronunciation modes are only the ones which correspond to cases where two opposite restrictions counterbalance each other.

As far as the pronunciation of *siglae* is concerned, Portuguese is in many aspects close to French. However, it presents some significant differencies that reflect distinctive parametrizations. The most evident case is precisely the *siglae* with a CVV structure which, in Portuguese, are not preferentially spelled. Contrarily to what happens with French, syllable nuclei may branch and, therefore, some VV sequences are interpreted as diphthongs (e.g. *FAO*). The two vowels of the VV sequence, however, may be in a hiatus, without necessarily triggering spelling. This is the case of *CIA*, for instance, whose segmental sequence is very common in word final position, where the post-stressed hiatus is well tolerated. The only CVV *siglae* which are systematically spelled are the ones that contain two identical vowels, a situation that does not occur in the common lexicon.

Phonological restrictions are important for the acceptability of a given segmental sequence as a "word" of the language, but what is at stake is not its acceptability as a "possible word", but rather as a "probable word". Certain *siglae* such as *AR*, for instance, which are homographs of a common lexicon word, are obviously possible words, though they are always spelled as a company (or public service) name. However, the frequence of occurrence of monossylabic words in the lexicon is extremely reduced (if the weight of functional words is ignored), and those which have an empty onset are even less frequent than the others. In fact, one can notice that the set of features that may be used to explain the pronunciation mode of *siglae* also explains the frequence of occurrence of words with the same prosodic structure in the lexicon. Languages in which empty nuclei are allowed may present syllables whose vowel is not phonetically realised.

They can also interpret non-syllabic consonants in a sequence as onsets or codas of syllables of this type. This is the reason for reading some *siglae* with sequences of obstruents which would not be syllabified otherwise. Empty constituents, however, are always marked structures which, in some sense, inhibit the reading of *siglae* and whose effects seem to be cumulative: the inhibition of reading is always higher for a *sigla* with two empty constituents than for a *sigla* with a single one. *Siglae* with a VC$\mathring{\text{v}}$ pattern, such as *IPO*, are more frequently spelled than others with a VCV or CVC pattern, for instance. Although consecutive empty nuclei in word final position are possible in Portuguese (e.g. *síntese*, most often pronounced as ['sĩ.t.z]), *siglae* with a CVCC pattern in which the CC are obstruents are generally spelled (e.g. *RATP*). The same is not true when the two consonants are syllabified and the *siglae* only contains one empty nuclei in absolute word final position (e.g. *SERB*).

A small set of rules was designed to take into account most of the above restrictions. This set was used to predict the pronunciation mode of *siglae* in the Portuguese lexicon. In 95% of the cases, the results were in agreement with the option taken by the manual transcribers.

5 NATIVISED PRONUNCIATIONS OF FOREIGN NAMES

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-phone 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.

5.1 Factors influencing nativization

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.

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.

5.2 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 [10], 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.

5.3 Context

The pronunciation of foreign names 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 ?".

5.4 Narrowing the phonetic transcription

The placement of syllabification marks for names of foreign origin also posed some interesting problems, although it was not mandatory for the inter-language lexicon. 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 criteria. Syllabification

errors due to lack of knowledge of the foreign language morphology may occur, independently of the criteria used for the native language.

6 CONCLUSION

The paper presented several issues that concern the pronunciation of proper names and described the two major linguistic resources produced by the ONOMASTICA project. The potential of self-learning methods for grapheme-to-phone conversion was discussed and so were the problems raised by acronyms and by the nativisation of foreign names. 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.

ACKNOWLEDGEMENTS

The ONOMASTICA project was a joint effort involving many researchers from different Universities whom the author gratefully acknowledges. This paper therefore mentions contributions from 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), Andreas Mengel (Institut für Fernmeldetechnik), Peter Molbaek (CPK), Georg Ottensen (Sintef Delab), Jose Pardo (UPM), Vito Pirrelli (ICL), Mark Schmidt (CCIR), Andrew Sutherland (CCIR), Francisco Valverde (UPM), and François Yvon (Telecom Paris). We would particularly like to acknowledge the contribution of our colleagues at INESC and CLUL: Fernando Silva and Isabel Mascarenhas.

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.

[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.

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

[6] V. Pirrelli and S. Federici, "You'd better say nothing than say something wrong: analogy, accuracy and text-to-speech applications", Proc. of the European Conf. on Speech Technology, Madrid, 1995.

[7] M. Plénat, "Observations sur le mot minimal en Français". In Laks & Plénat (eds), De Natura Sonorum, Presses Universitaires de Vincennes.

[8] 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.

[9] T. Sejnowski and T. Rosenberg, "Parallel networks that learn to pronounce English text", Complex Systems, 1987.

[10] 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.

[11] A. Villalva, "Compounding in Portuguese", Rivista di Linguistica, Vol. 4, no. 1, 1992.