ON THE PRONUNCIATION MODE OF ACRONYMS IN SEVERAL EUROPEAN LANGUAGES

I. Trancoso and M. C. Viana
INESC/IST CLUL
INESC, R. Alves Redol, 9, 1000 Lisbon, Portugal.
Tel. +351 1 3100268, FAX: +351 1 3145843, E-mail: Isabel.Trancoso@inesc.pt

ABSTRACT

The paper describes our research work concerning the pronunciation mode of acronyms in German, French, and Portuguese. Most of the rules are related with the well-formedness of the constituents and the minimum and maximum weight thresholds required for reading and spelling an acronym.

The results of the tests for the three languages were considered very promising, reaching decision errors below 4%. The rule set was also applied to a very small English corpus, with relative success. We believe that further optimisation is still possible, if language specific parametrisation is taken into account, in particular for the languages where a limited corpus of acronyms was available.

1. INTRODUCTION

This paper describes the work that we have been doing on the pronunciation of acronyms, first in the scope of the European LRE project ONOMASTICA (Multi-language pronunciation dictionary of proper names and place names), where we have mainly concentrated on acronyms in our own language (Portuguese), and more recently in the scope of the European Telematics project VODIS (Advanced Speech Technologies for Voice Operated Driver Information Systems), where the two working languages are French and German.

The pronunciation of acronyms raises interesting problems not only because different pronunciation modes can be used (some acronyms are read, some are spelled), but also because when read, acronyms frequently deviate from the grapheme-to-phone rules derived for the common lexicon. Moreover, in some languages, the pronunciation of acronyms shows considerable variability even among native speakers.

These features motivated a study of the lexical processes involved in the formation of acronyms and of their constituent elements which is summarised in the following section. We shall then proceed to describe the training and testing corpora used in this work. Next, basic generic rules for deriving the pronunciation mode of acronyms will be briefly outlined for 4 European languages together with the corresponding results obtained with our corpora. Before discussing future research directions we shall present results with an interesting set of acronyms of radio stations in French and German.

2. FORMATION AND PRONUNCIATION OF ACRONYMS

Several types of lexical formation processes can be found in corpora of acronyms: acronymy (in the strict sense), blending, *siglae* (or abbreviation proper) and, although rarely, single truncation. In most cases, company acronyms involve abbreviations of the general designation of the company or of one or more names / surnames of it(s) 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 important distinction between acronyms in the strict sense and *siglae* is not, however, 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.

Another type of lexical formation process which can be often found in acronyms is compounding - root compounding or word compounding. Whereas acronyms of the first type are typically written as single words, acronyms of the latter type are typically written as separate words (often with hyphens or spaces). In languages like Portuguese, for instance, the difficulty of distinguishing between the two types of compounding is one of the sources of the variability of pronunciation of acronyms by native speakers, as different strategies of vowel raising may apply to the two types.

The pronunciation of acronyms has been the subject of many recent research studies. Most of these studies concern the two different pronunciation modes of acronyms (reading vs. spelling). In some Grapheme-to-Phone systems, spelling seems to be triggered by the existence of dots in between letters. This rule, however, is not applied by native speakers when pronouncing an acronym with dots. The most basic rule for deciding whether one acronym should be read or spelled is to spell acronyms without vowels and read the remaining ones. This rule is implemented in most GtoP systems but fails in too many cases. In fact the condition of having at least one vowel is necessary for reading, but is not sufficient. The length of the sequence is an important

feature in the reading/spelling decision: with rare exceptions acronyms with less than 3 characters are spelled and the ones which have 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, and others such as CVVV are always spelled.

This type of observations has been made for several languages and has been on the basis of some attempts to explain the pronunciation mode as a function of the interaction of different prosodic constraints. Hence, for instance, in order to be read, each segmental sequence must allow a syllabic parsing in agreement with the set of general principles and with language specific parametrisation. However, it must also correspond to a possible word pattern, both in extension and weight. In some cases, the structural and weight constraints may not be compatible and the resolution of this conflict depends on their relative importance. Plénat [2] proposes minimum and maximum weight thresholds for a *sigla* in French to be read, and refers some examples of possible conflicts. The minimum threshold of two *morae* (corresponding to a monossylable with branching rhyme or to a disyllable) defines a limit below which a *sigla* is mandatorily spelled, and the maximum threshold of three syllables defines another limit above which *siglae* are preferably read. These syllabic weight constraints coexist with a set of structural ones which determine the spelling of *siglae* whose constituents may be considered ill-formed.

The fact that the pronunciation mode for *siglae* can vary from language to language and that this variation may be interpreted in terms of its specific parametrisation emphasises the importance of phonological constraints in the acceptability of a given segmental sequence as a "probable" word of the language.

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 why some *siglae* may be read, although they contain sequences of obstruents which would not be syllabified otherwise.

Other recent studies are not only devoted to the pronunciation mode of *siglae* but rather to their reading strategies which may significantly differ from the one used for the common lexicon. For French, for instance, Yvon [4] found significant differences in vowel nasalisation, in the pronunciation of the letter "e" and of vowel groups, and in the status of final obstruents.

According to this author, when reading *siglae*, people seem to choose among all the possible realisations the ones that allow the listener to better reconstruct the graphical form, while respecting the usual reading rules. It seems likely that this constraint of maximising the information conveyed to the listener may also play a role in the choice between reading and spelling. If no realisation is found that simultaneously satisfies the usual reading rules and the constraint of maximum information, given that all other factors are equally satisfied, than the *sigla* will be most likely spelled. In order to illustrate this point, Yvon takes the example of the acronyms CEC and CIC which are almost always spelled in French, whereas SEC and SIC are more often read.

3. TRAINING AND TEST CORPORA

Our main source of acronyms for deriving pronunciation rules and testing them has been the Onomastica pronunciation lexicon. In spite of totalling 11 languages, not all of them include company names. Even for those languages which include company names, the ones that are acronyms are not always marked or distinguishable by their orthography (with capital letters, for instance).

Our study so far has been restricted to Portuguese, French, German and English. Norwegian is currently in progress. For some languages we had access to the full company names. For others, we had only isolated words. We restricted our study to company names formed by a single word and, for the sake of simplicity, we have excluded names with digits, as elementary rules may be easily applied to pronounce the digit part in them. We have likewise excluded acronyms with plus signs or ampersand signs linking two words (or letters), as specific rules can be used to pronounce them. Acronyms with mixed pronunciations (partly read and partly spelled) were also excluded in this first approach, since they can only be dealt with using morphological parsing in order to detect frequently used elements (e.g. *soft* in *GSoft*).

The dimension of the sub-corpora of acronyms significantly differs from language to language. For Portuguese, we have identified 26,603 acronyms. For French we had 14,870 acronyms. For German, we had only 740 acronyms. This short corpus was complemented by another one provided by Robert Grudszus, of Robert Bosch GmbH. It includes company names extracted from a telephone CD for the whole Germany. The original list included 1200 acronyms (some of them also present in the original Onomastica corpus) and an indication of whether they were either read or spelled. The total German corpus had 1767 acronyms. For each language, approximately 70% of each sub-corpus was used for training the rule set, and the remaining 30% for testing it. For English, we could only identify in the Onomastica lexicon the acronyms which were spelled. This amounts to only 114 acronyms which were used as a test set for commonly used rules.

4. RULES FOR DECIDING THE PRONUNCIATION MODE

As a first step towards deriving rule sets for the pronunciation mode of acronyms in the 3 first languages, acronyms were parsed into syllabic constituents (onsets, rhymes and nuclei), assuming that those are maximally binary branching. Only general sequencing constraints concerning the sonority and complexity of adjacent elements were considered. Most of the rules we shall describe are related with the well-formedness of the constituents and the minimum and maximum weight thresholds described above.

The first general rule we have applied consists of reading all acronyms with diacritics. These were found in the Portuguese and German training corpora (7% for the first and less than 1% for the latter). Diacritics are mainly used for stress assignment or for modifying the vowel quality, functions that do not make sense in spelled acronyms. Consonants with diacritics are also frequent in some languages, but rarely in the beginning of a word. Hence, they do not contribute to the formation of *siglae*.

The second general rule for deciding the pronunciation mode of acronyms consisted of spelling all the acronyms with no vowels. In fact, a consonant-only acronym cannot be parsed into syllabic constituents. This rule covers a significant percentage of some of the training corpora (40% for French, 2% for Portuguese, 27% for German and 31% for the English test set).

The third rule was not applied consistently to all languages. It consists of spelling all acronyms with less than 3 characters and reading the ones with more than 5. These length constraints are closely related with the minimum weight threshold for reading proposed by Plénat, which is not reached by 2-character acronyms. On the other hand, spelling acronyms with more than 5 characters would often correspond to more than 5 syllables, a length that is not favoured. In French, however, we have found around 30 acronyms of length greater then 5 which were marked as spelled in the training corpus, so this part of the rule was not consistently applied.

The fourth rule consists of spelling acronyms with certain consonant clusters (e.g. CCC$, C$CCC, CCC$, CC$C, etc., where the dollar sign represents the syllabic boundary mark). The length of the consonant cluster depends on its position within the acronym (beginning, middle, end) and the length of the acronym (e.g., a 4-character acronym with a CVCC structure will be spelled, whereas a longer acronym with the same ending pattern may be read). In spite of the fact that the syllabification criteria we have adopted admits empty nuclei, those are not considered well-formed constituents, if they are not licenced by a following vowel. The parsing into consecutive onsets implies a sequence of violations of phonotactic constraints which leads to spelling. The dependence on the position is related to the different acceptance of empty nuclei at the left or right edges of an acronym, the latter being more easily tolerated.

The fifth rule consists of spelling all acronyms which start by a vowel and have no other vowels, that is, they start with a syllable with no onset, followed by a sequence of empty nuclei syllables, thus implying consecutive violations. This rule covered a significant number of the remaining acronyms for French and Portuguese, and very few for German.

The sixth rule consists of spelling all acronyms with no consonants, since they also correspond to consecutive ill-formed syllables. This rule yielded some 15 exceptions for Portuguese, and a couple for the other languages.

The seventh rule consists of spelling some 4-letter and 3-letter patterns (e.g. CCV, CCV, VVC). These patterns either do not fulfil the minimum weight requirements for reading or they imply consecutive violations. The applicability of this rule varied from language to language, mainly due to the different dimensions and constitution of the acronym corpora available for each language.

The two following rules were not applied to all languages either. The first one consisted of spelling all acronyms with a repeated letter in the beginning. And the second one consisted of spelling 3-letter acronyms which end in “e”. In fact, this final vowel is often interpreted as a schwa and hence may or may not be pronounced. Therefore, the identification of the orthographic form of the acronym would be ambiguous, if it was read.

Most of the acronyms not covered by this large set of rules are read acronyms. In fact, a large majority includes CV$CV or VC$CV patterns, thus fulfilling the minimum threshold requirements for reading. Although further constraints could be imposed to cover the spelled acronyms that do not include these patterns, we considered their coverage too limited for the size of our corpus.

5. RESULTS

This section summarises the results obtained for the training and test sets. For the French training set an error percentage of 1.5% was obtained. This corresponds to 148 acronyms which should be spelled and are read according to the rules and to 12 acronyms which should be read and are spelled. A similar percentage was obtained for the test set. For the Portuguese training set, an error percentage of 0.6% was obtained. This corresponds to 31 acronyms which should be spelled and to 83 acronyms which should be read. Similar results

were obtained for the test set. For the German training set an error percentage of 3.3% was obtained. This corresponds to 14 acronyms which should be spelled and to 27 acronyms which should be read. A lower percentage was obtained for the test set (1.8%). For the English test set, 18 acronyms were found which should be spelled and are read (16%).

It should be noted that most of the exceptions to the derived rules correspond to acceptable pronunciations. In many cases, very similarly structured acronyms can be found which are also pronounced as these exceptions. The error percentage has decreased as the dimension of the training set increases, which was expected. The different constitution of the corpora was evident in the results. The total French and German corpora had a much higher percentage of spelled acronyms (around 50%) than Portuguese (4%). The relative small number of spelled acronyms for Portuguese resulted in more exceptions that should be read and are spelled, than vice-versa. The total number of rules for spelling was therefore larger for French, somewhat lower for German and Portuguese, and significantly shorter for English, due to the relative small size of the corpus.

6. TESTS WITH ACRONYMS OF RADIO STATIONS

An interesting test of the derived pronunciation mode rules for German and French was made using the corpus of acronyms of radio stations built in these two languages in the scope of the VODIS project. In this project, in fact, speech input is used to command an equipment called Berlin, which among other tasks such as navigation, GSM dialling and CD player control, also enables switching from one radio station to another. For both languages, the following sub-corpora were found (examples in German and French, respectively):

Read:
ANTENNE, BROCKEN, RADIOTON
CULTURE, NOSTALGI, VOLTAGE
Spelled:
AFN, DLF, FNN, SWF
BFN, RFM, TSF
Read or spelled with digits:
BREMEN1, RADIO5, BFBS1
EUROPE1, RTL2
Mixed pronunciation modes
(partly read and partly spelled):
RPR ZWEI, RMBRADIO (RMB Radio)
CHERIEFM (Cherie FM), OUIFM (Oui FM)
Always expanded:
D-RADIO (Deutschland Radio)
STADTRA (Stadt Radio)
N.DAME (Notre Dame)
T EIFFEL (Tour Eiffel)
Frequently expanded:
NDR (Norddeutscher Rundfunk)
WDR (Westdeutscher Rundfunk)
LATINA (Radio Latina)
MUSIQUE (France Musique)
Plays with words:
N-JOY (Enjoy)
NRJ (Energie)
Foreign words:
LIFE
SKYROCK (Sky Rock)

For the read / spelled acronyms, the pronunciation mode was always correctly determined by the rule set.

7. CONCLUSIONS AND FUTURE DEVELOPMENTS

Three different sets of rules were derived to determine the pronunciation mode of acronyms for French, Portuguese and German. The results of the tests for these languages and also for English were considered very promising, although we believe that further optimisation is still possible, if language specific parametrisation is taken into account, in particular for the languages where a limited corpus of acronyms was available.

Most of the rules are related with the well-formedness of the constituents and the minimum and maximum weight thresholds. There seems to be a close relationship between the number of allowed violations and the number of well-formed constituents in the acronym.

Future research efforts should be invested namely on the study of how native subjects pronounce acronyms which are read, and how this pronunciation differs from the one predicted by GtoP software used for the common lexicon.

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the VODIS project Robert Grudszus and Arian van Hessen for their collaboration in this work.

8. REFERENCES

[1] P. Mareüil, "*Vers la Phonématisation automatiques des sigles*", La Linguistique, Vol. 31, no. 1, 1995.

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

[3] 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.

[4] F. Yvon, "*Régles de Transcription Graphéme-Phonéme pour la prononciation automatique des sigles*", Lynx, 30, 1994.