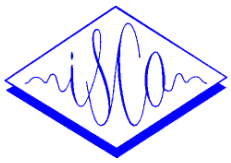
ISCA Archive
http://www.isca-speech.org/archive



3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

AUTOMATIC GRAPHEME-TO-PHONEME CONVERSION OF DUTCH NAMES

Emmy M. Konst and Louis Boves

Department of Language and Speech, Nijmegen University
P.O. Box 9103 6500 HD Nijmegen, The Netherlands
E-mail: Konst@LETT.KUN.NL

ABSTRACT

This paper describes the work on grapheme-to-phoneme conversion for Dutch proper names, carried out in the framework of the ONOMASTICA project. We give some basic statistics of Dutch names, describe the development of a set of letter-to-sound rules optimized for proper names, discuss the main reasons why a specific rule set for this purpose is necessary and finally give results of two tests of the performance of the optimized rule set.

I. INTRODUCTION

A growing number of speech technology applications deal with names. Correct pronunciation of names is essential in applications like 'reverse directory inquiry' and 'who is calling'. Names often cause problems in speech synthesis systems using rule-based grapheme-to-phoneme conversion. The rule set, designed for letter-to-sound correspondence for 'regular' words, is normally not suited for name pronunciation [1,2,3]. In Dutch (as in many other European languages) the problems are mainly due to two causes: (1) immigration, i.e., the occurrence of many names of foreign origin and (2) spelling conservatism, i.e., the spelling of person names does not follow spelling reforms for regular words.

Even when letter-to-sound rules are adapted for names, a certain percentage of the names will still be mispronounced. The use of a pronunciation lexicon, containing high quality, hand checked pronunciations can solve this problem. In this respect, of course, names do not behave differently from regular words in a language like Dutch, which is rich in loan words. However, it must be stressed that, due to continuing immigration, a lexicon of phonemic forms does not obviate the need for letter-to-sound rules optimized for names.

ONOMASTICA, a project in the 'European Commission Framework Programme Linguistic Research and Engineering', will make available a large pronunciation lexicon for European names from 11 different languages in machine readable form [7]. In addition, ONOMASTICA aims to develop a set of grapheme-to-phoneme rules for names in each of these languages. This rule set is used to create a provisional pronunciation, which is then checked and if necessary corrected by a competent phonetician. As more hand-checked grapheme-phoneme correspondences become available, this information can be used to update and improve the rules.

In this paper we describe the development of grapheme-to-phoneme rules for native Dutch last names in some detail. Furthermore, we present the results of a performance test. We will start with detailed statistics of first-, last-, street-, and topographical names.

A separate set of rules has been developed for first names; these rules cannot be tested in a formal way, because all available names have been used for rule development. Work is under way to estimate the language background of non-native family names; for each major group a separate set of rules is being developed.

II. STATISTICS OF NAMES

Statistics of person names occurring in the Netherlands are based on the Dutch PTT Telecom telephone subscribers data base. In the Netherlands Register Offices are decentralized, and cannot be combined to obtain a national registry; this makes the Telecom subscribers data base by far the most comprehensive single source of person last names in the country. This data base comprises 245,517 different last names. Almost half of these names (117,130) occur only once. We suspect that a fair proportion of the unique names are due to misspellings: quite a number of orthographic forms look like a misspelling of a name that in its correct spelling is (much) more frequent. For the other part these unique forms pertain to foreign names; here too, spelling problems abound: the alphabetic transcriptions of names from languages using non-Latin characters or having a phonemic system that is very different from Dutch may be inconsistent, adding to the number of unique or infrequent forms.

In the recent past it was unusual to have first names mentioned in the telephone directory, but a growing number of people nowadays have their first names listed. The first names occurring in the Telecom data base were merged with names from the Spectrum First Name Book [4], constituting a list of 27,464 first names. Unfortunately no frequency information is available in [4], but as first names are strongly affected by fashion it is obvious that only a limited set will have a real high fre-

Table 1: Total number of unique names per category and number of names occurring only once (singlets).

Type	total #	singlets
last names	245,517	117,130
first names	27,464	not known
place names	2,353	not known
street names	117,246	24,737

quency of occurrence. Moreover, there may be considerable discrepancy between first names as occurring in official documents and the names used in practice. This is an additional argument for not trying to obtain statistics about first names. Obviously, a large number of first names is of foreign origin.
Place names are also taken from the Telecom data base. The large majority of major communities have unique names; there is a fair number of small communities with non-unique names, but there is no generally accepted register of names of hamlets, locations, etc. Therefore, it is not possible to state reliably how many place names are really unique. Place names constitute the smallest set of names.
Street names are extracted from the Postal Code table maintained by PTT Post; this table is guaranteed to be complete, even for the near future. A summary of the name statistics is given in Table 1.

III. GRAPHEME-TO-PHONEME CONVERSION OF NAMES

It has already been said that an automatic grapheme-to-phoneme conversion program can accelerate the creation of a pronunciation dictionary, if it produces sufficiently accurate phoneme representations. Our basic automatic grapheme-to-phoneme conversion program for Dutch (FONPARS1) is rule-based [5]. In our text-to-speech system these rules are used to handle words which cannot be found in the pronunciation dictionary. FONPARS1 employs context sensitive rewrite rules, which need to be strictly ordered. A number of different modules can be distinguished in the grapheme-to-phoneme rule set for regular words, the most important of which comprise rules for spelling normalisation, morphologic analysis, letter-to-sound conversion rules, syllabification and stress assignment rules.
In order to create an optimal rule set for Dutch names most modules in the system for regular Dutch words had to be adapted. The adaptation process will be described in some detail below.

3.1 Stress assignment

The stress assignment rules implemented in our text-to-speech system were not adequate for names. In Dutch monomorphemic arbitrary words main stress is always located on one of the three final syllables, depending on the relative weight of their syllable-rimes [6]. Dutch names do not violate this rule, but seem to have a preference for non-final stress: in most person names, and especially in last names, the first syllable takes the stress. In a sample of 45,000 native last names from our data base 89.6% of the names bear stress on the first syllable, 6.0% bear final stress, and 4.4% bear penultimate stress. (It should be mentioned that this sample did not merely comprise monomorphemic names.)
In order to make the stress assignment rules more suited for names, rules assigning final stress were deleted or restricted to cases needed for names. Moreover, a new rule was added to the end of the stress assignment module. This rule by default stresses the first syllable of polysyllabic names that have not yet been provided with a stress mark. Obviously, this final rule only works on syllables that can take stress.

3.2 Morphological analysis and letter-to- sound rules

Grapheme-to-phoneme correspondence was a more complicated problem to deal with. One specific problem with names is caused by the frequent use of ancient, often redundant spellings. For example, the diphthong /ɛɪ/ which has already two different spellings in modern Dutch ("ei" or "ij", depending on the underlying Old Germanic vowel) is often spelled "eij" in names. Thus the name "Arkesteijn" corresponds to "Arkestein" or "Arkestijn" in modern spelling. Other examples are the use of "ae" for "aa" as in the name "Adriaens", or the use of "aau" for "au" in the name "Blaauw". The spelling of Dutch has gone through a number of formal reforms since the middle of the 19th century, which have removed these ancient forms from regular words. Very few families have decided to go through the hassle (and the expenses) of adapting the spelling of their names to the new rules. The letter-to-sound rules designed for regular Dutch words fail on names with ancient spelling, e.g. because graphotactically illegal vowel sequences like "aau" are parsed as containing a syllable boundary.
Most of these problems were solved by enlarging the set of spelling normalization rules, which form a grapheme-to-grapheme conversion module. This module rewrites ancient grapheme combinations into their modern equivalent.
Furthermore, as is also the case in regular words, the grapheme "e" causes many problems. In Dutch "e" can be pronounced as [e], [ɛ] and [ə], mainly depending on morphological structure. Both compounding and derivation cause problems here, and both are different for regular words and names. Problems with compound analysis are aggravated by the fact that many morphemes which occur in names have disappeared from regular words. Moreover, names contain other prefixes and suffixes than

words. For example, for regular words the suffixes "-tje, -tjes" suffice, while the list of similar suffixes in names is much longer: "-tje, -tjes, -tjen, -tjer, -tjens, -tjers".

3.3 Syllabification

Syllabification in names does not deviate from syllabification in Dutch regular words. Nevertheless some modifications were needed, because the output of the rules was unsatisfactory in names. This concerns especially cases where suffixes are incorrectly separated from a name, as in names ending in "stra" or "stein".

IV. RULES ASSESSMENT

After having adapted the text-to-speech rules into an optimal set of rules for names, the performance of this new rule set was tested. To check whether the adaptations were real improvements, the performance of this set was compared to the performance of the original rule set for Dutch words implemented in the text-to-speech system. A test set of 1000 native last names randomly selected from names with frequency of occurrence between 12 and 15 was used. None of these names had been involved in the rule development process, so they were all "unseen". To distinguish native (actually: Germanic) from non-native names a list of 109,921 entries, provided by the German ONOMASTICA partner, containing information on the probable origin of the names was used as a first filter. The remaining names were checked by hand, and all forms which clearly have a non-Germanic origin were removed.
For each name in the test set was checked whether the output transcription of the rules was correct compared to the manual transcription of the name. Subsequently errors were recorded and categorized in the following error categories:

- transcription errors (including substitutions, deletions and insertions)
- syllabification errors
- stress assignment errors

4.1 Multiple pronunciations

A small part of the names in the test data base can have multiple pronunciations. The grapheme-to-phoneme rules, however, produce only one pronunciation for each name. If the output of the grapheme-to-phoneme rules matched one of the possible pronunciations, it was considered correct, even if it was not the phonetician's first choice. (In the eventual lexicon the preferred pronunciation is, of course, added).

Table 2: Number of errors made by rules for Dutch names compared with rules for Dutch words, classified in three error categories.

error type	rules for Dutch names	rules for Dutch words
transcription	78	573
syllabification	20	160
stress assignment	75	281

V. RESULTS

In this section the results of the performance test are described and discussed.

5.1 Rule set for names

Assessment of the rule set on the 1000 unseen native last names yields a correct performance of 87.2%. In 128 names a total of 173 errors are made, distributed over the error categories as shown in Table 2.
With 78 mistakes transcription errors constitute by far the largest category. Though much attention has been payed to improvement of the conversion of the grapheme "e" in the new rule set, 46 of the transcription errors (more than 50%) relate to this grapheme. Failure to obtain correct decomposition of ancient compounds which are no longer recognized as such are an important cause of the trouble.
The category of syllabification errors comprises 20 errors. Eight of these are caused by an error in phonemic transcription.
Stress assignment errors are in 18% of the cases (n=14) related to transcription errors; the remaining errors are true stress assignment errors.

5.2 Rule set for Dutch words

The grapheme-to-phoneme rules for normal Dutch words applied to the same test set give only 45.0% correct conversion. In 550 names a total of 1014 errors are made (see Table 2).
Again most errors are transcription errors: 573 errors in transcription of phonemes were found. Many names contain more than one transcription error, mainly due to wrongly inserted phonemes, caused by the ancient spelling of the name. Incorrect conversion of the grapheme "e" is another frequent transcription error.
The number of syllabification errors made by the rules for Dutch words is much larger than the errors made by the adapted rule set for names. Rules for Dutch words produce such large number of syllabification errors partly because of wrong phoneme insertions, resulting in extra syllables. The remaining errors in this category are true errors in syllabification.

Stress assignment is incorrect for 281 names in the test set. Stress assignment errors are related to wrongly inserted phonemes, as well as to wrong conversion of the grapheme "e", and to inappropriate stress assignment rules.

5.3 Test set size

A test set of 1000 names may perhaps be considered as rather small. We could not increase the size of the independent test set due to the time-consuming nature of detailed human error analysis. To investigate whether the set was a representative sample of Dutch last names, we performed another test on 47,880 native Dutch names, selected from the hand-corrected lexicon. This set includes more frequent names, but also names with a low frequency of occurrence (down to 6). To simplify the test procedure, only one phonemic form per name was retrieved from the lexicon. The overall performance of the rule set for names on this large test set is 87.7%. This result is essentially equal to the outcome of the test with the independent test set. Thus, the set of 1000 names used in the detailed test yielded a representative test result.

VI. CONCLUSION

Many errors are made by letter-to-sound rules for regular Dutch words, when tested on a set of names: only 45.0% of the names is correctly converted by these rules. The rule set designed for names gives a much higher percentage of correct pronunciation, namely 87.2%. Apart from a few remaining conversion problems for the grapheme "e" we can state that the rule adaptation has been quite successful. Additional analysis of the errors related to this grapheme might lead to further improvement of the rules. We should not forget, however, that correct phonemic mapping of the "e" depends largely on the morphological structure of a name, which cannot always be computed. Finally a considerable number of errors seems to stem from 'true' exceptions, proving the need for a dictionary based approach.

References

[1] B. Van Coile, S. Leys and L. Mortier. "On the development of a name pronunciation system." *ICSLP 92*, Vol 1, pp. 487-490, 1992.

[2] R. Carlson, B. Granström and A. Lindström. "Predicting name pronunciation for a reverse directory service." *Eurospeech-89*, Vol 1, pp. 113-116, 1989.

[3] K. Belhoula. "Rule-based grapheme-to-phoneme conversion of names." *Eurospeech-93*, Vol 2, pp. 881-884, 1993.

[4] J. van der Schaar, D. Gerritzen and J.B. Berns. *Spectrum Voornamenboek.* Utrecht: Het Spectrum. 1992.

[5] J. Kerkhoff, J. Wester and L. Boves. "A compiler for implementing the linguistic phase of a text-to-speech conversion system". *In: Linguistics in the Netherlands*, H. Bennis and W. van Lessen Kloeke (eds), pp. 111-117, 1984.

[6] R.W.J. Kager. *A metrical theory of stress and destressing in English and Dutch.* Dordrecht: ICG Printing. 1989.

[7] M.S. Schmidt, S. Fitt, C. Scott and M.A. Jack. "Phonetic transcription standards for European names (ONOMASTICA)." *Eurospeech-93*, Vol. 1, 279-282, 1993.