ONOMASTICA - Creating a Multi-Lingual Dictionary of European Names

Joakim Gustafson, Department of Speech Communication and Music Acoustics, KTH

ABSTRACT

The objective of the ONOMASTICA project is to make available quality controlled pronunciation lexicons of European names in machine readable form (CD-ROM). Transcription of up to 1.000.000 names per language will be produced in a semi-automatic way. We at KTH are responsible for the Swedish part of the project, in co-operation with TELIA who provided us with the name database. This paper will present the work we have done so far and discuss some of the problems we have met.

INTRODUCTION

The ONOMASTICA project was established as part of the "European Commission Framework Programme - Linguistic Research and Engineering", with the goal to provide:

- a multi-language pronunciation dictionary
- letter-to-sound rules
- statistics on names, their frequencies and inter-occurrences
- self-learning software

A total of eleven languages are covered in the project: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish.

The 2-year project will not only produce pronunciation dictionaries for up to one million names per language, but will also investigate the problems of exchanging national names amongst the partners to create a matrix of "nativised" pronunciation of foreign names in each language.

The multi-lingual dictionary could for example be used in the following sectors:

- Telecommunications: Automated directory enquiry systems, telephone banking and enhanced talking newspapers and books for the blind.
- Consumer sector: Map information and guidance systems, talking dictionaries and courseware systems for pronunciation teaching.
- Publishing: Hard-copy as well as electronic dictionaries containing pronunciation.

The ultimate pronunciation dictionary would include a carefully verified manual transcription of each name, but due to the limited resources only a subset of the name list can be transcribed and verified manually. The names will be divided into three different Quality Bands defined as:

BAND I: Transcriptions judged to be **correct** to the best of a competent phonetician's knowledge. Transcriptions are guaranteed correct for some owners of the name.

BAND II: Transcriptions judged by a competent phonetician to be **acceptable** to a native speaker/listener. Names that cannot be easily verified.

BAND III: Transcriptions not yet checked by a competent phonetician. Names that have been transcribed automatically.

The names in BAND I & II will be chosen according to their frequency in the telephone directory so that a cumulative coverage of 80% will be obtained. Each partner has transcribed a "Golden Set" of 40,000 names as part of the project's first phase.

TRANSCRIPTION

We are to transcribe personal names, company names, street names and town names, where the company names have been ignored because they are so many and only occur once. Automatic transcription would be very difficult since company names consist of a mix of invented words, foreign words and acronyms.

The place names have been manually transcribed with the help of Garlén (1991). We have transcribed the 3100 places included in this book according to our conventions of transcription.

The street names often consist of ordinary Swedish words followed by "gatan"(street) or "vägen"(road), ex. Blåmesvägen (Blue tit road). They are quite easy to transcribe if you use a Two-Level morphology (Koskenniemi 1983), with a large lexicon. Our Swedish lexicon consists of 40,000 morphs (Karlsson 1990).

Christian names and other first names can be difficult to transcribe since they may have a foreign origin that influences the pronunciation. But fortunately only about 500 names need to be transcribed to obtain a coverage of 80%. However, we have transcribed 6000 names, obtaining a coverage of 95%. The names that can be used in double-names (Lars-Erik, Anna-Lotta) have been included in the lexicon (920 female and 680 male first names), so that these forms can be generated automatically. The double-names have a very general pronunciation structure: the male are usually three-syllabic and have accent I with stress on the second syllable. The female vary more, but are usually three or four-syllabic and have accent II with primary stress on the first syllable (Kvillerud 1980).

Finally, we have the surnames. Since they have been our primary concern, they will be described in more detail.

SURNAMES

Until recently we have worked with a name database derived from the Greater Stockholm telephone directory, but we now work with one from the whole of Sweden. The work done with the Stockholm database will be reusable in the future work.

When transcribing the surnames automatically the following steps are taken: First, the origin of the names is determined. Then they are processed by different sets of rules depending upon their origin. The process is described in Figure 1.

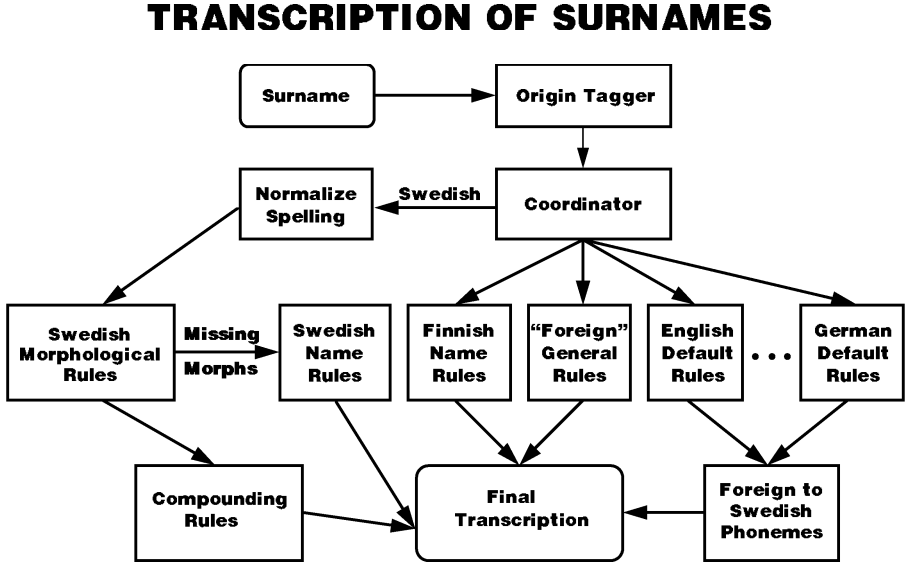

Figure 1. *Flow chart on transcription of names*

Determining the origin of the names

It is easier to transcribe a name if its origin is known. Since the system is designed to imitate a Swedish person attempting to pronounce a foreign name, it is not certain that the origin-tags will be etymologically unequivocal. However, the goal is that they should make the same decisions about language origin as people with ordinary language knowledge do. To date 23 tags for origin have been included. The choice of these has been arbitrary so a revision will be necessary.

The tagging is done using the KTH text-to-speech system with phonological rules that recognise patterns that are specific for different languages. Since 80% of the names in the database are Swedish, it is a good approximation to consider all names Swedish, until a pattern specific to another language is found.(Carlson, Granström, Lindström, 1990)
We have also run some tests with neural nets, but did not achieve better results than with rules. Origin tagging with neural nets requires a large tagged database. With such a database for training, a better result may be achieved with this approach.

Normalising spelling of names

People with ordinary names sometimes make their names more unusual and interesting by spelling them in an unorthodox way. To address this situation we received a list from TELIA with different spellings of the same names. These different spellings do not always follow any orthographic conventions. One practice that has been observed is the insertion of "h", e.g., Bhlom, Gusthafson, another the use of "x" instead of "ks" in son-names, e.g., Ericxson, Erixcon. Some other popular replacements are:
s→ z, k→q, k→c, å→aa, ä→ae, ö→oe, ö→eu, o→ou, i→ie, f→ph, v→fv, v→w, j→i.
The spelling of the names must be normalized in order to simplify the automatic transcription.

Morphological structure of surnames

Swedish surnames can be divided into three groups:

I. Those that combine a male first name with the suffix -**son**
II. Those that are compounds of two morphs ex Berg#ström(mountain#stream)
III. Others

The first two groups are very suitable for a morphological analysis. In our lexicon we have so far included the surname morphs described in Table 1.

Table 1. *Description of surname morphs incuded in lexicon*

Group of morphs	Example	Number of morphs
full name entries	svensson	2650
male first names that combine with son	anders-	680
compound forming morphs for surnames	-berg-	1084
stress-taking morphs for surnames	-elius	44
non-stress taking morphs for surnames	-ner	87

In order to obtain a better coverage, but not overgenerate, these morphs and the rules for combining them still need adjustment.

The third group will be generated by the last two non-compound forming groups of morphs, but since these morphs are so few, not all of these names will be found with our two-level morphology program. These remaining names will be run through ordinary Swedish letter-to-sound rules adjusted for these names. We have not yet sufficiently studied which these names are and hence the adjustments that need to be done.

Problems with transcription of foreign names

The ability to pronounce foreign names is dependent on the person's competence in reading and pronunciation (Mengel 1993). These competencies can be defined as:

Reading competence - The ability to recognise and identify foreign orthographic name structures.

Pronunciation competence - The ability to produce sound structures of foreign names. This ability could be due to the speaker's knowledge in the foreign language or to the fact that his sound system is comparable to that of the foreign language.

If these abilities are set to + or -, and the possibility of an incompetent listener who is able to produce correct pronunciations is neglected, three different ways to pronounce a foreign name is obtained, as shown in Table 2.

Table 2. *Possible pronunciations of foreign names correlated to competence of speaker.*

Native Reader	Reading Competence	Pronunciation Competence	Pronunciation of Jones ['ʒoʷnz]
I	**-**	**-**	**[ˋju:nəs]**
II	**+**	**-**	**['jou:ns]**
III	**+**	**+**	**['ʒoʷnz]**

I The native speaker does not recognise the orthographic pattern as foreign because of lack of knowledge, therefore he cannot produce the correct pronunciation.

II The speaker recognises the pattern as foreign but cannot pronounce it adequately, thus adapts it to his native phonetic pattern.

III The educated expert who recognises the pattern and pronounces it adequately.

The first two strategies are possible to generate with most text-to-speech systems. To generate the third strategy the missing foreign phonemes have to be added, possibly including rules for realisations. The question we face is to what extent foreign phonemes should be incorporated, and whether to incorporate different acoustic realisations of the same phonemes.

The work that is being done within this project will not only be useful in different applications as mentioned initially, but also in cross language studies.

ACKNOWLEDGEMENTS

This research has been supported by a grant from NUTEK.

REFERENCES

Carlson, R. Granström, B. Lindström, A. (1990): "Automatic Generation of Name Pronounciation for a Reverse Dictionary Service," Final report, Dept. of Speech Communacations and Music Acoustic, KTH.

Garlén, C. (1991): "Svenska ortnamn uttal och stavning"

Karlsson, F. (1990): "SWETOL: A Comprehensive Morphological Analyser For Swedish", manuscript, Dept. of General Linguistics.

Koskenniemi, K. (1983): "Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production" Dept. of General Linguistics, University of Helsinki.

Kvillerud, R. (1980) "Förnamn i Göteborg- Namnskick för skolbarn födda 1958".

Mengel A. (1993): “Transcribing Names - Multiple Choise Task: Mistakes, Pitfalls and Escape Routes.” Onomastica Research Colloquium, pp. 5-9.