TRANSCRIBING NAMES WITH FOREIGN ORIGIN IN THE ONOMASTICA PROJECT

Joakim Gustafson

Department of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

This paper studies the problem of transcribing foreign names. The transcriptions of first names in five languages have been studied to show examples of how this problem has been dealt with in the Onomastica Multi-Lingual Pronunciation Dictionary of European names.

The paper describes this dictionary and the methods used to do the automatic transcriptions for the Swedish part.

INTRODUCTION

Names have a different morphology and phonology compared to ordinary words. This is the reason why the normal letter-to-sound rules used in general text-to-speech systems are inadequate for the transcription of proper names. To deal with the name pronunciation problem, name transcription rules and a name dictionary have to be developed. The objective of the Onomastica project is to produce such rules and a dictionary of European names that will be published on a CD-ROM. This paper will present the problems encountered in the work on this project, and how these have been solved. The transcriptions of first names in five languages are examined to illustrate the problem. The Swedish name transcription system will be presented as well.

THE ONOMASTICA DATABASE

The objective of the ONOMASTICA project, funded by the LRE-programme, is to build a quality controlled, multi-lingual pronunciation dictionary of proper names in Europe. The project covers eleven languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. Transcription of up to 1.000.000 names per language will be produced in a semi-automatic way.

The ultimate pronunciation dictionary should include a carefully verified 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 are transcribed in three different quality bands, where the first band includes transcriptions judged to be correct for some owners of the name. The second band gives transcriptions that are acceptable to a native speaker/listener. The third band contains names that have been transcribed automatically, without manual checking. The names in bands I & II were chosen according to their frequency in the telephone directory, so that a cumulative coverage of at least 80% was obtained.

The Swedish database, see Table 1, consists of the whole Swedish telephone directory, containing 4.5 million subscribers. The names that occurred more than five times were selected for transcription in band I, obtaining a cumulative coverage from close to 95 % for surnames to 100% for place names (almost all places have more than five subscribers).

Table 1. The Swedish Name Database

Name category	# of names	names with frequency >5
Surnames	228048	46859
Place names	6373	6120
Titles	27055	5370
Street names	65196	39822
First names	60850	10479

THE TRANSCRIPTION SYSTEM

The existing KTH text-to-speech system has been modified and upgraded to cope with proper names. (See Figure 1.) [3]. First the origin of the name is determined to simplify the work for the automatic transcriber [5]. 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 correct. However, the goal is that they should make the same decisions about language origin as people with ordinary language knowledge would do. To date, 23 tags for origin have been included. The tagging is done using the KTH text-to-speech system with phonological rules that recognise patterns that are specific to different languages [2].

Depending on the origin, each name is sent to a different set of grapheme-to-phoneme modules. The Swedish names are first sent through a TwoLevel-morphology analyser (TWOL) [4] with general Swedish morphs augmented with 1200 name-morphs and a name-lexicon with names occurring in the Stockholm telephone directory compiled during a previous project [2]. The morphology approach is especially suitable for names in Sweden because they are often multi-morphemic. From the morphology analyser morphs with stress and boundary markers are obtained. A set of phonological rules merges these into complete transcriptions. The names that were not transcribed by TWOL are processed by the ordinary Swedish letter-to-sound rules adjusted for names. The foreign names are first run through language-specific letter-to-sound rules with language specific phonemes. These phonemes are then mapped to the closest Swedish equivalents.

All names were manually corrected by the same person in order to obtain consistency. Different tools were used in this process ranging from UNIX scripts to the KTH text-to-speech system. The method of correcting the transcriptions using both orthography, transcription and synthesised speech has proven to be both fast and efficient [3].

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 problem we received a list from the Swedish PTT with different spellings of the same names.

The use of different spellings seems to be more popular in Sweden than in the other four languages examined in this paper. In Swedish only 81% of the first names have a single spelling compared to 97% in Italian, where a sequence of names is used to make the name unique. Swedish first names have up to 24 different spellings, Italian names have up to 6. The different spellings do not always follow ordinary orthographic conventions. One practice that has been observed is the insertion of "h", Bhlom, another the use of "x" instead of "ks" in names ending with "son", for example, Ericxson. Some other popular replacements are: s→z, k→q, k→c, å→aa, ö→oe, ö→eu, i→ie, f→ph, v→fv, v→w.

The spelling of the names must be normalised in order to simplify the automatic transcription.

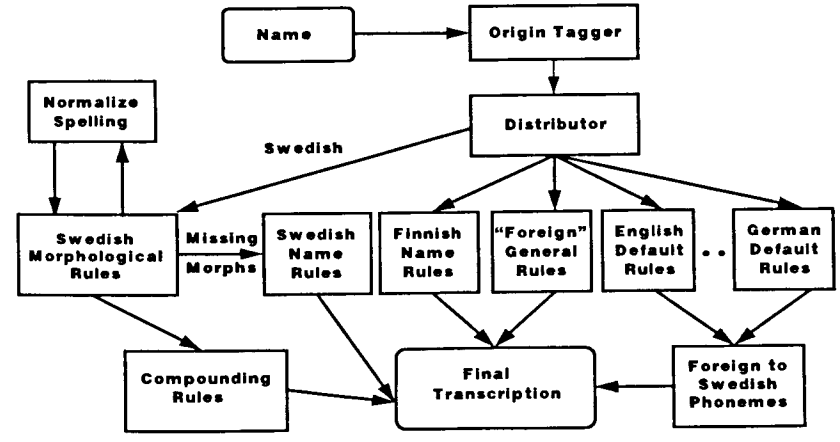

Figure 1. The KTH-system for transcription of names.

A STUDY OF THE FIRST NAMES IN FIVE LANGUAGES

To exemplify the problems of transcribing names with foreign origin the databases containing first names from Great Britain, France, Germany, Italy and Sweden were examined. These varied in size from 10000 in Sweden to 35000 in Italy. (See Table 2)

Table 2. The number of transcribed first names in the five databases.

Sw	En	Fr	Ge	It
10461	16111	12383	31979	35013

The difference in size of the databases could be adjusted by selecting the 10.000 most frequent names, but since the frequency only was available in the Swedish database the databases could not be truncated. All the transcribed names are used in the study.

The structure of names differs from common words, since names often move with people across borders, and adjust to the new language. Table 3 shows the mean number of letters and phonemes in the first names and in the 10.000 most frequent common words [1].

Table 3. Mean number of letters and phonemes in First Names (FN) and Common Words (CW) in the languages.

	letters in FN	letters in CW	phonemes in FN	phonemes in CW
Sw	7.4	7.4	5.6	6.9
En	7.0	7.1	5.6	6.0
Fr	8.9	7.6	6.3	5.2
Ge	8.1	8.7	6.2	7.8
It	10.7	7.4	9	6.9

The names were transcribed in different phonetic alphabets with broad transcriptions. To be able to compare the transcriptions done in the different languages, they were converted from the various phonetic alphabets to IPA. But since broad transcriptions were used the actual realisation of individual phonetic symbols will vary from language to language.

The most common phonemes in each language's first names are shown in Table 4. The table shows that the most common phoneme is [a] in all languages, except for English where it is the [ə].

Table 4. The most common phonemes in the transcriptions of first names, with percentage figures to the right.

Sw		En		Fr		Ge		It	
a	12	ə	9	a	13	a	10	a	15
n	7	ɪ	8	i	10	t	7	o	11
r	7	n	8	ʀ	8	n	7	e	9
l	7	æ	7	l	7	r	6	i	8
ɪ	7	r	6	n	6	l	6	n	8
s	6	l	5	ɛ	5	ɪ	5	r	8
e	6	i	4	m	5	ə	5	l	6
t	4	s	4	s	4	k	4	t	5
m	4	ɛ	4	d	4	ɐ	4	m	4
k	4	m	4	t	4	s	4	j	4

In all languages, except Italian, the ten most common phonemes cover about 60% of all occurring phonemes. In Italian they cover 77%. Italian has the least number of phonemes (28) but the largest number of phonemes per name (9). Swedish and English, however, have the largest number of phonemes (about 40), but the smallest number of phonemes per name (about 5.5). If you pick the names from each country that contains as many as possible of these phonemes you get the following names:

Sw Nils-Einar [nɪlsejnar]
En Alexander [ælɪgzɑndər]
Fr Alexandrine-Marthe [alɛksãdrinmaʀt]
Ge Weichselgärtner [vaɪksəlgɛrtnɐ]
It Vittorio-Emanuele [vitorjoemanuele]

The databases altogether contain 88.000 different names. 79.000 of these only occurred in one country, 981 occurred in all five. The length and stress markers were removed from the transcription of these common 981 first names and the transcriptions were compared. Table 5 shows that the most similar languages are Swedish-German and French-Italian, and those that are most dissimilar are German-Italian. In Italian foreign names often get an Italian spelling, for example Jesus is spelled Gesu in Italy.

Table 5. Number of names that get the same transcription in the language-pairs

	Sw	En	Fr	Ge	It
Sw	-	115	121	201	113
En	115	-	116	115	102
Fr	121	116	-	102	193
Ge	201	115	102	-	87
It	113	102	193	87	-

Table 6. The pronunciation of an initial J in first names, number of occurrences. The likely origins of the names are indicated within the parentheses.

Swedish	English	French	German	Italian
j 458	dʒ 642	ʒ 858	j 558	j 188
ɧ 31 (Fr,Sp)	j 50 (Sw,Ge)	dʒ 37 (En)	dʒ 29 (En)	dʒ 79 (En)
ʂ 12 (Fr)	ʒ 12 (Fr)	j 14 (Sw,Ge)	ʒ 24 (Fr)	i 19 (Fr)
dj 3 (En)	h 6 (Sp)	x 9 (Sp)	x 4 (Sp)	x 17 (Sp)

When examining the databases it was noticed that the letter J in initial position got quite different transcriptions in the five languages (See Table 6). The different ways it can be pronounced seem to be dependent of the likely origin of the name. The names that are considered to be of a certain origin get the pronunciation of "J" that is most common in that language, or is mapped to the closest one in the native language. English names are mostly transcribed with [j] in Swedish, but in some cases the [dj] have been used to imitate the English [dʒ]. Spanish names like Juan [xwan] has been transcribed with the same phonemes in German, French and Italian, but [ɧuan] in Swedish and [hwan] in English.

CONCLUSIONS

The transcription of foreign names presents some problems. There are a number of factors that influence the realisation of a foreign name:

- the level of education in foreign languages
- the phoneme inventory and prosody of the foreign name is frequently adapted to the language spoken
- the context in which it is produced, such as the receiver of the message.

The work on ONOMASTICA has shown that there are a number of decisions that have to be made, such as:

Q1 How to transcribe a foreign name if you don't know the origin

A1 Use the same pronunciation rules for foreign names as for native:
The Swedish name Greger ['gre:gər] gets the Dutch transcription ['xrexər].

Q2 How to deal with foreign phonemes that do not exist in the native language.

A2 a) Map the foreign phonemes to the closest native:
The English name Winston ['wɪnstən] is transcribed ['vɪnstɔn] in Swedish.

A2 b) Enlarge the native phoneme inventory:
The English phonemes [ð] and [θ] are added to the Swedish inventory to get Heather['hɛðər] and Keith['ki:θ].

Q3 How to deal with foreign graphemes.

A3 a) If the realisation of the grapheme in the foreign language is known use the closest native phoneme:
The Swedish town Göteborg [jø:tebɔrj], is transcribed [jɛtɛbɔrg] in Greek, where the Swedish way to pronounce "ö" is known, but it has to be mapped to the closest Greek phoneme.

A3 b) If the realisation of the grapheme is unknown map it to the closest native grapheme:
Göteborg is mapped to Goteborg in Spanish and it is transcribed [gɔtɛβɔr].

ACKNOWLEDGEMENT

The work on the Swedish part of the Onomastica project has been supported by grants from NUTEK.

REFERENCES

[1] Carlson, R. Elenius, K. Granström, B. Hunnicutt,S (1985): "Phonetic and orthographic properties of the basic vocabulary of five European languages" STL-QPSR 1/1985 pp 63
[2] Carlson, R. Granström, B. Lindström, A (1990): "Automatic generation of name pronounciation for a reverse dictionary service," Report, Dept. of Speech Com. Music Ac , KTH.
[3] Gustafson, J.(1994): "Onomastica - Creating a multi-lingual dictionary of European names", w. papers 43, Lund Univ. Dept. Ling. pp 66-70.
[4] Koskenniemi, K (1983) "TwoLevel Morphology: A general computational model for word form recognition and production" Dept. of General Ling. , University of Helsinki
[5] Vitale, T (1991):"An algorithm for high accuracy name pronunciation by parametric speech synthesizer" Computational Linguistics, Vol. 17, No. 3, pp.257-76