USING TWO-LEVEL MORPHOLOGY TO TRANSCRIBE SWEDISH NAMES

Joakim Gustafson

Email: joakim_g@speech.kth.se
Department of Speech Communication and Music Acoustics,
KTH, Stockholm, Sweden

ABSTRACT

Names are difficult to handle for normal letter-to-sound rules, since these usually are designed for ordinary words. The structure of Swedish names differ from ordinary words - but their multi-morphemic structure make them suitable to analyse with a morphological analyser.

The paper presents the work on names from the Swedish telephone directory, as part of the ONOMASTICA project [7], including a brief study of the structure of Swedish names.

The speech communication group at KTH have developed a system where a morphology analyser is used together with a set of rules to transcribe ordinary Swedish words. This paper will describe the work done to extend this system to cope with names as well.

The paper shows that the approach of transcribing Swedish names with the Two-level Morphology analyser (TWOL) is appropriate.

INTRODUCTION

Names are different in their structure compared to ordinary words, and because of this 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 pronunciation dictionary of 8.5 million European names, that will be published on a CD-ROM.

This paper will describe the structure of Swedish names and how a morphology analyser can be used to transcribe them.

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 band I includes transcriptions judged to be correct for some owners of the name. Band II gives transcriptions that are acceptable to a native speaker/listener. Band III 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 is shown in Table 1. It 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 STRUCTURE OF SWEDISH NAMES

Names are difficult to transcribe since they do not have a unique spelling. Some of these spellings are invented by people with common names, who want to make their names more unusual by spelling them in an unorthodox way. The use of different spellings seems to be more popular in Sweden than in other languages. 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 26 different spellings (Ann-Christine), while Italian names only have up to 6 spellings. 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”, the name Eriksson, for example, has a spelling 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.

Names also have a different morphology and phonology compared to ordinary words, which makes them difficult to handle for ordinary letter-to-sound rules. Proper names can be difficult to transcribe since they may have a foreign origin that influences the pronunciation.

Names in Sweden are often multi-morphemic; street names for example generally consist of one or two common words followed by -gatan (street) or -vägen (road), for example **Blåmesvägen (Blue tit road)**. Table 2 shows some statistics.

Table 2. Some statistics on Swedish names and common words.

	Mean number of letters	**Mean number of phonemes**	**Mean num. of syllables**	**Consonant /Vowel ratio**
Surnames	7.3	6.6	2.4	1.8
Street names	11.3	13	4.3	1.8
First names	7.4	5.6	2.4	1.3
Place names	9.4	8.7	2.9	1.7
Common Words	7.4	6.9	2.7	1.6

First names are often combined into double names. These have a uniform pronunciation structure, but the structure of male and female double names differ; male double names are usually three-syllabic and have accent I with stress on the second syllable **(Karl-Axel [ka:lʾɑksəl]).** The female double names vary more, but are usually three or four-syllabic and have accent II with primary stress on the first syllable **(Anna-Lena [ʾʾɑnɑ#lʾe:nɑ])**. To handle this difference in pronunciation first names are tagged with sex in the morph lexicon

The 10,458 first names that were transcribed in band I were processed with the morphology analyser, and 2,538 got the male tag and 2,728 got the female. These tagged names cover 97% of the occurrences in the Swedish telephone directory. Their stress patterns are shown in table 3. In the first column the stress pattern of the name is described with two numbers, x:y, where x is the number of syllables and y is the syllable that has the main stress.

Table 3. The stress pattern of Swedish first names.

Stress pattern	**Male names**		**Female names**	
	%	**Example**	**%**	**Example**
1:1	7	Bo	3	Ann
2:1	35	Arne	41	Eva
2:2	3	Renee	4	Marie
3:1	5	Mikael	20	Annika
3:2	43	Johannes	11	Agneta
3:3	1	Severin	4	Marianne
4:1	0		11	Märta-Stina

Double names are common among both the male first names (50%) and the female first names (30%), but many of these constructions have low frequency. To account for this the names were weighted by frequency when measuring the numbers of syllables. The male first names got a mean number of syllables of 1.8 and the female 2.7.

The female names are almost one syllable longer than male, as have been found in earlier studies [5]. For both male and female names the stress pattern 2:1 is very common. Single-syllable names are more common for men than for women, and four-syllable names only occur for women.

Swedish surnames are often multi-morphemic as well. They have a very uniform structure, that can be divided into three main groups:

I. Names that combine a male first name with the suffix -son**,** ex**.** **Gustavsson (the son of Gustav)**

II. Names that are compounds of two root-morphs, often nature related, ex. **Ek#ström (Oak# stream)**.
The first morph is one of 4016 while the second is one of only 610.

III. Others, ex. **Lanner**

The 46,856 transcribed surnames were examined and the result is displayed in table 4. Two and three syllable surnames are most common in Sweden. In this material compounds of surnames such as Lundblad-Dabrovski were split, which explains why the names with more than four syllables are mostly foreign. Almost all surnames with accent II (99.1%) have their main stress in the first syllable. For surnames with accent I the position of the primary stress vary more, but 65% of them are on the first and 30% on the second syllable.

Table 4. The structure of surnames occurring in Sweden.

#	**1**	**2**	**3**	**4**	**5**	**6**
Number of names with # syllables	3,193	25,867	15,059	2,390	320 mostly foreign	21 only foreign
Accent II compound names with primary stress in syllable #	18,393	50	1			
Accent II-names with primary stress in syllable #	5,082	132	24			
Accent I-names with primary stress in syllable #	15,100	6,983	971	100	9 only foreign	

MORPHOLOGICAL ANALYSIS OF SWEDISH NAMES

The Swedish morphology analyser SWETWOL was constructed by Fred Karlsson in 1988-89 [3] based on the two-level morphology TWOL, designed by Kimmo Koskenniemi [4]. SWETWOL is based on “classical” Swedish grammar and can form words by inflection, derivation and compounding. The SWETWOL analyser consists of a lexicon with more than 45,000 vocabulary items, where the bulk of the words where derived from Svenska Akademiens Ordlista (SAOL-10,11) and a set of eight two-level rules that are compiled into run-time finite-state automata. These rules are used when the lexical and the phonemic surface representations differ.

All the morphs in the dictionary have been transcribed manually based on an automatic procedure, using the KTH text-to-speech system [6]. The transcriptions of all morphs in a word are included among the parts of speech tags.

The system has been updated to cope with names as well. The current morphological lexicon, listed in table 5, consists of the 45,000 general Swedish morphs augmented with 2,623 transcribed name-morphs and a name-lexicon with transcriptions of names occurring in the Stockholm telephone directory, compiled during a previous project [1]. Most of the work has been done on the surnames, since they often are made of more than one morph, taken from a set of surname morphs.

Table 5 The updated Swedish morphological lexicon.

Group of morphs	Example	# of morphs
ordinary Swedish morphs	hopp-	45,000
place names	stockholm	4,361
full surnames	olson	24,083
first names, uncategorized	afsaneh	1,570
female first names	-marie-	2,243
male first names	-anders-	1,931
initial only, compound forming surname morphs	wahl-	1,877
compound forming surname morphs	-berg-	554
stress-taking surname morphs	-elius	72
non-stress taking surname morphs	-ner	120

From the study of the 46,856 transcribed surnames described previously, it can be found that there are 2,324 VCV-structures in surnames. Most of these names are accent I (1,157), and the accent II names are either single morphemic (307) or compounds (860). Names that are compounds consist of two parts were the first part is one of 4016 possible morphs. The second part is taken from a set of 610 morphs. Some morphs can only be in either of these two positions, which leads to a total number of 4,237 compound forming surname morphs. Many of these morphs have been constructed by adding **-en, -er ,-e** or **-s** at the end of a morph. This structure has been implemented in the morph lexicon by allowing some morphs to be “inflected” with these endings. This has reduced the number of root-morphs necessary in the lexicon.

The same morphs that are used in compounds are often used in non-compound names as well, but with different endings. There are about 150 endings that are used to generate these names, for example: **-lert, -man, -ing** and **-ner**. There are 72 morphs in the lexicon that always have final position and that get the primary stress regardless of the first part. Ex **-lander,** **-in** and **-elius**. This structure has been implemented in the same way as derivation for common words.

All morphs that have been included in the morph lexicon have restrictions on their use according to the names the actually exists in the Swedish telephone directory. Only morphs that occur with the -er, -en, -e endings are allowed to use those, even though some other theoretically could be used with them. Name morphs that were found only in initial position are not allowed in any other position. This has been done to ensure that the system generates names that follow the same conventions as people in Sweden do when they create names. The morphs only cover correctly spelled Swedish names, which explains why the coverage, shown in figure 1, decreases with rank. The last 90,000 surnames only occurred once in the database and most of these where either foreign or misspelled.

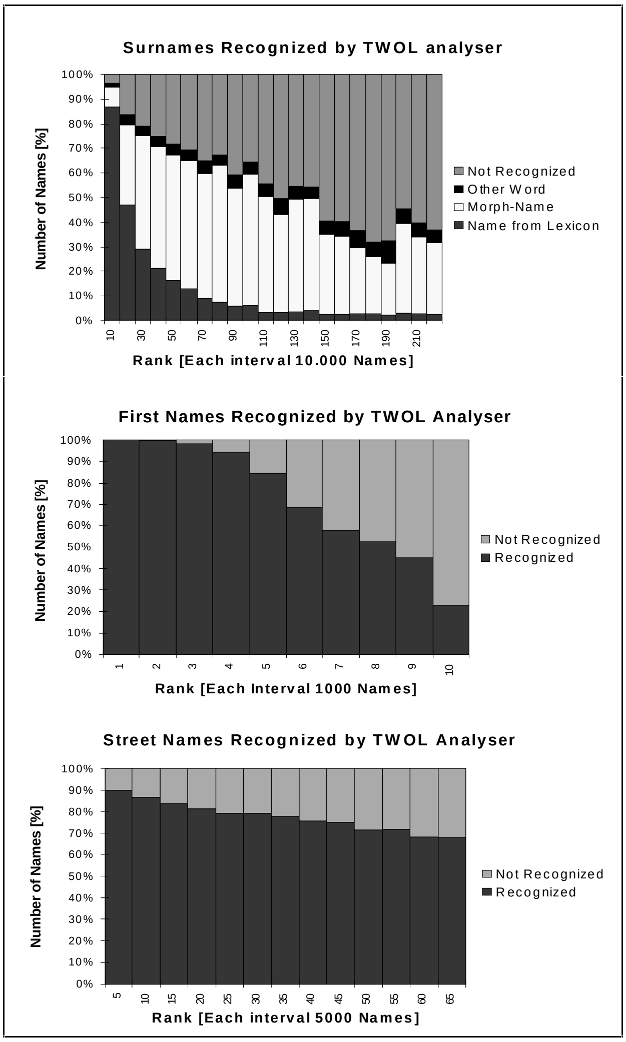

Figure 1. The coverage of the TWOL analyser.

THE TRANSCRIPTION SYSTEM

The KTH text-to-speech system upgraded to cope with proper names has the following structure:

First the origin of the name is determined to simplify the work for the automatic transcriber. It is not certain that the origin tags will be etymologically correct, since the system is designed to imitate a Swedish person attempting to pronounce a foreign name. 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 rules that recognise patterns that are specific to different languages.

Depending on the origin, each name is sent to different grapheme-to-phoneme modules. The Swedish names are first sent through the morphology analyser, from which morphs with stress- and boundary-markers are obtained. A set of phonological rules merges these into complete transcriptions. The names that cannot be formed by these morphs are processed by Swedish letter-to-sound rules 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.

RESULTS

All names were automatically transcribed and manually corrected by the same person in order to obtain consistency. Different tools were used during correction, ranging from unix-scripts to the KTH text-to-speech system. The method of correcting the transcriptions using both orthography, transcription and synthesized speech has proven to be both fast and efficient [2].

To evaluate the system a test set of 1000 names was randomly selected from the first 30,000 surnames. The total error rate for these names was 4%. About 90% of the words were processed by the morphology analyser giving only 2.4% errors while the letter-to-sound gave 20% errors. More than 50% of the names that were not processed by the morphology were of foreign origin. To be able to study the quality of the transcriptions that had been produced with the morphology approach a new test set was selected. A thousand names that could be formed by the name-morphs were randomly selected from the rank interval 30-40,000. These names were processed by either TWOL, letter-to-sound rules for names or letter-to-sound rules for common words. The three approaches produced transcriptions with the following error rates:

the letter-to-sound-rules for common words	**66%**
the letter-to-sound-rules for names	**52%**
the TWOL approach	**7%**

The errors produced by TWOL was mostly of morphological nature, such as missing morph boundary, or missing stress in end-morph, giving wrong accent in the transcription. Most of these errors can be fixed by tuning the morph lexicon.

The quality of all the transcriptions in the project has also been measured by an audit of independent auditors who were native speakers of respectively language [8]. A total of 1000 names from each quality band was presented to the auditor and the transcriptions were examined. This quality test showed that the Swedish band I transcriptions had an error rate of 0.3%.

The band I names were those that occurred more than five times in the Swedish telephone directory. To increase the cumulative coverage for the surnames a second set was selected to be transcribed in band II. These surnames were selected among those that occurred five times or less. The names were first tagged automatically, then those that were tagged as Swedish were run through the TWOL analyser. The ones that could be formed by TWOL were selected, which gave 75,000 automatically transcribed names. The test described earlier where the TWOL approach had an error rate of 7% included names not tagged as Swedish. And not all of the 7% ‘not correct’ transcriptions were wrong. Most of them were considered to be possible pronunciations, acceptable for a native speaker. It was consequently considered safe to put these transcriptions in band II without checking. The result from the audit shows that this was OK, since no wrong transcriptions at all were found among the 1000 in the test sample.

The band III names were also checked and the error rate was 53.3%. The 194,000 low frequency names in band III were run through TWOL and/or letter-to-sound rules. These names were mostly foreign names or misspelled names, which could explain the higher error rate for these names.

This paper has shown that a morphology approach for transcription of names in Sweden is quite effective, especially for names with Swedish morphs.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Carlson, R. Granström, B. Lindström, A (1990): "Automatic generation of name pronounciation for a reverse dictionary service," Report, Dept. of Speech Com. Mus.KTH.

[2] Gustafson, J.(1994): “Onomastica - Creating a multi-lingual dictionary of European names”, working. papers 43, Lund Univ. Department. Linguistics pp 66-70.

[3] Karlsson, F. (1990): “A comprehensive Morphological Analyzer for Swedish“, Dept. of Gen. Ling. Univ. of Helsinki.

[4] Koskenniemi, K (1983) "TwoLevel Morphology: A general computational model for word form recognition and production", Dept. of General Ling., University of Helsinki

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

[6] Magnuson, T. Granström, B. Carlson, R. Karlsson, F. (1990):"Phonetic transcription of a Swedish morphological analyzer," in Proc.of Fonetik-90, Phonum 1, Reports from the Dept. of Phonetics Univ. of Umeå.

[7] “ONOMASTICA Multi-Language Pronuciation Dictionary of Proper names and Place Names”, Technical and Financial Annex, Project No. LRE-61004

[8] “ONOMASTICA Multi-Language Pronunciation Dictionary of Proper names and Place Names”, Final Report,Project No. LRE-61004