

Abstract

Names are common in most text-to-speech applications, such as automatic reading of newspapers, reverse directory services and voice communication aids. The letter-to-sound rules included in these applications often cannot handle names, since the rules usually are designed for ordinary words. The structure of Swedish names differs from ordinary Swedish words - but their multi-morphemic structure makes them suitable to analyse with a morphological analyser. A study of the structure of Swedish names in the Swedish telephone directory is presented in this thesis.

The thesis also presents the work done within the Swedish part of a European Linguistics Research and Engineering project, called the Onomastica project. The aim of this project was to produce a multi-lingual pronunciation dictionary for names occurring in 11 European languages, as well as developing techniques for automatic transcription of names.

The speech communication group at KTH has developed a system where a morphology analyser is used together with a set of context dependent rules to transcribe ordinary Swedish words. This thesis describes the work done to extend this system to cope with names as well. Other transcription methods are also described. Finally the system is evaluated showing that the approach of transcribing Swedish names with the two-level morphology analyser is efficient.

CONTENTS

1. Introduction.....	1
2. The Onomastica project.....	3
3. The Swedish name database.....	7
4. The structure of Swedish names.....	11
4.1. Surnames.....	11
4.2. First names.....	15
4.3. Place names.....	20
4.4. Street names.....	21
5. Grapheme-to-phoneme conversion.....	25
5.1. Context dependent rules.....	25
5.2. Symbolic learning.....	26
5.3. Artificial Neural networks.....	27
5.4. Markov Models.....	27
5.5. Analogy.....	28
5.6. Morphological analyser.....	30
5.7. The Onomastica results.....	33
5.8. Summary.....	33
6. Description of the transcription system.....	35
6.1. Grapheme-to-phoneme conversion rules.....	36
6.2. Morphological analysis.....	37
6.3. Normalising the spelling of names.....	40
7. Transcribing names with foreign origin.....	41
7.1. How to deal with foreign names.....	41
7.2. The origin tagger.....	42
7.3. Comparison of first names in five languages.....	44
7.4. Pronunciation of an initial ěJí in different languages.....	45
8. Evaluation of the name pronunciation system.....	49
8.1. Evaluation of three test samples.....	49
8.2. The Onomastica audit evaluation.....	56
9. Conclusions.....	57
10. Acknowledgements.....	59
11. References.....	61
12. Appendix.....	65

1. Introduction

Speech synthesis is used in many information systems today. One of the most used systems with speech synthesis is the reversed directory service in the United States. In this telephone service the name and address of the subscriber with a certain telephone number are read by a speech synthesiser.

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 procedures and a name dictionary have to be developed.

To provide the European Community with a name pronunciation resource a European Linguistic Research and Engineering project was established, called the Onomastica project. This project has produced transcription techniques and a pronunciation dictionary of 4.5 million European names. The dictionary that was produced within the Onomastica project has been published on a CD-ROM.

This thesis will present an analysis of the structures of Swedish names and compare them with the structure of non-name words in Swedish. Different techniques used to transcribe names will be described, concentrating on how a morphological analyser can be used in a transcription system for names.

There are a number of factors that influence the pronunciation of names. The spelling of the names do not follow the same conventions as ordinary words do. This could be because old spellings remain in names or because people spell their names in unusual ways. One of the most difficult factors to handle in the transcription of names is the origin of the name. This thesis will give some examples of how foreign names can be handled in a name transcription system.

The main part of the thesis is focussed on the description and evaluation of the Swedish name transcription system. The thesis will show that a morphological analyser is an excellent tool when transcribing Swedish names.

2. The Onomastica project

Many applications in language engineering require automatically obtained correct pronunciation of names. Examples of Text-to-Speech applications that include names are:

Reading books and newspapers Our Swedish lexicon of 122,000 words derived from 143 million words of running text from books and newspaper contains 15,000 names.

Reverse directory services The application of reading the telephone directory, deals mainly with names.

Road Guidance Systems The driver gets information about the best way to drive, and news, like traffic jams or roads under construction (Includes many place names).

Voice Communication Aids The TTS-system is used to produce speech for people with speech handicaps. Names are of course an important part of the communication.

The automatic transcription of names is a difficult task that requires a large dictionary and/or grapheme-to-phoneme conversion techniques that can cope with names as well as common words. It is almost impossible to produce a dictionary that is large enough since there are so many names. In the US there are about 1.5 million uniquely spelled names (Social Security Administration, 1985). The Onomastica project was established to provide the EU community with a quality controlled, multi-lingual pronunciation dictionary for up to 1 million names per language in Europe. A total of eleven languages were covered in the project: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish. A dictionary can never cover all names, since new names are introduced all the time, for example by immigrants, therefore appropriate grapheme-to-phoneme conversion techniques for names have to be developed. The Onomastica project was part of the "European Commission Framework Programme - Linguistic Research and Engineering" (LRE). The project lasted from 1 January 1993 to 30 June 1995, to the cost of 3,637 kECU of which 1,500 kECU was funded by EU. The goal of the project was to provide:

- a multi-language pronunciation dictionary
- grapheme-to-phoneme conversion techniques for names
- statistics on names, their frequencies and inter-occurrences
- self-learning software

The Onomastica project included 11 Academic Partners, who developed the grapheme-to-phoneme conversion techniques and transcribed the names. The project also included 11 associated partners. These were national telephone companies, that provided machine readable name files as background intellectual property to the project. All 22 partners in the Onomastica project are listed in Table 1.

Table 1. The partners in the Onomastica project.

COUNTRY	Academic Partner	Associated Partner
Great Britain	CCIR, University of Edinburgh	BT Laboratories, Martlesham
Denmark	CPK, University of Aalborg	Jydsk Telefon, Aarhus
France	ENST, Paris	France Telecom (CNET), Lannion Cedex
Germany	Institut für Fernmeldetechnik, Berlin	Deutsche Bundespost Telekom, Darmstadt
Greece	Department of Electrical Engineering, Patras	Intrasoft, Athens
Italy	Inst. Of Comp. Linguistics, Pisa	CSELT, Turin
Netherlands	Department of Language and Speech, Nijmegen	PTT Research, Leidschendam
Portugal	INESC, Lisbon	Telefones de Lisboa e Porto, Lisbon
Spain	UPM, Madrid	Telefónica, Madrid
Norway	SINTEF DELAB, Trondheim	Norwegian Telecom Research, Kjeller
Sweden	Dept. of Speech, Music and Hearing, KTH, Stockholm	Telia Promotor, Solna

The ultimate pronunciation dictionary would include a carefully verified transcription of each name, but due to the limited resources only a subset of the name list was transcribed and verified manually. The names were 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, due to limited resources.
- BAND III:** Transcriptions not yet checked by a competent phonetician.
Names that have been transcribed automatically.

The names in BAND I & II were chosen according to their frequency in the telephone directory so that a cumulative coverage of at least 80% was obtained. Each partner transcribed a "Golden Set" of 20-50,000 names as part of the project's first phase. These were used to develop grapheme-to-phoneme conversion techniques.

The Onomastica project has produced a CD-ROM dictionary with names and their transcriptions. The Onomastica CD-ROM contains 4.5 million entries in Quality Band I & II. The content of the total project lexicon is shown in Table 2.

The complete Onomastica lexicon of 8.5 million names is available on EXABYTE tape format since the capacity of the CD-ROM used is inadequate for all of the lexical data. The CD-ROM and EXABYTE tape have been distributed to the 22 partners.

A subset of the lexicon can be obtained for research. Other organisations will be able to get subsets of the lexicon by contacting the telephone company that provided that part of the lexicon.

Table 2. *ONOMASTICA* lexicon entries for each language

Language	Quality Band I	Quality Band II	Quality Band III
English	137,721	159,417	800,000
Danish	135,503	0	169,935
French	185,761	8,870	773,172
German	947,316	316,792	632,949
Greek	1,433,847	0	0
Italian	84,847	69,909	596,459
Dutch	241,364	11,481	368,398
Portuguese	84,500	0	0
Spanish	188,760	133,787	59,700
Norwegian	73,329	2,198	472,898
Swedish	109,185	82,994	194,859
Inter-language	11,000 x 11 = 121,000	Not applicable	Not applicable
TOTALS	3,743,113	785,448	4,068,370

The names were transcribed by the Academic partners using their own machine readable phonetic alphabets in the accents specified in Table 3. The phonetic transcriptions on the CD-ROM have been translated from these different alphabets into the International Phonetics Association Standard Computer Coding (Esling, 1990).

Table 3. *Accents for each language in the ONOMASTICA lexicon*

LANGUAGE	ACCENT
Spanish	Castillian
Italian	Tuscan (Academic)
Danish	Copenhagen
German	Deutsche Hochlautung (Hanovarian)
English	Received Pronunciation (Rp)
Portuguese	Lisbon
French	Tours
Greek	Athens
Dutch	Standaard Nederlands (Sn)
Swedish	Standard Swedish
Norwegian	Oslo

The Onomastica project also investigated the problems of "nativised" pronunciation of foreign names in each language. All 11 partners in the project selected 1000 touristically interesting names, including cities, towns, airports, stations and places of

historic interest. These names were distributed to all other partners, for the purpose of making nativised pronunciations for each name. The result was a lexicon-matrix with 11,000 names transcribed in eleven different languages. This multi-lingual dictionary could for example be used in the following sectors:

- telecommunications, such as automated reversed directory services
- consumer sector, such as road guidance systems or talking dictionaries
- publishing, i.e. pronunciation dictionary of names
- linguistic research, to study how foreign names are transcribed in different languages.

The inter-language matrix could be useful in systems with speech recognition that are designed to be used by people of different origin.

3. The Swedish name database

The Swedish database, presented in Table 4, consists of the complete Swedish telephone directory, containing 4.5 million subscribers. The names that occurred more than five times were selected for transcription in band I. These names had a cumulative coverage from close to 91.2% for surnames, to 99.6% for place names. To increase the cumulative coverage for the surnames to 95% a second set was selected to be transcribed in band II. These surnames were selected from those that occurred five times or less. All names were first tagged automatically, from which two groups of names were selected. The first group contains names tagged as Swedish. These names were run through the morphological analyser. Names that could be formed by the morphological analyser were selected, which gave 75,000 automatically transcribed names. The second group are names tagged as Finnish. Many Finnish names are included in the Swedish database, and these are easy both to detect and to transcribe. About 7,500 names that were tagged as Finnish were consequently selected and transcribed.

Table 4. The Swedish Name Database.

Name category	Number of names	Names with frequency >5	Cumulative coverage with frequency >5
Surnames	228,048	46,859	91.2%
First names	60,850	10,479	98.4%
Street names	64,621	39,822	96.2%
Place names	6,373	6,120	99.6%
Titles	27,055	5,370	95.4%

The company names in the Swedish telephone book were not transcribed due to limited resources. There are 400,000 company names in the telephone directory, of very different importance. The problem is to select the most important for transcription. Company names are quite hard to transcribe for many reasons: the names are often invented by the owner, they often contain foreign words and many include acronyms. The structures of acronyms have been described by McCully and Holmes (1987).

A lexicon with 107,379 non-name words from a corpus of 150 million words, taken from newspapers and books, has been examined and will be used as reference in this thesis. The lexicon contains the most common of a total of 1.8 million word forms occurring in the corpus. Figure 1 shows the number of words in the different classes that is needed to obtain a certain cumulative coverage. The surnames have the same occurrence patterns as non-name words (NNW).

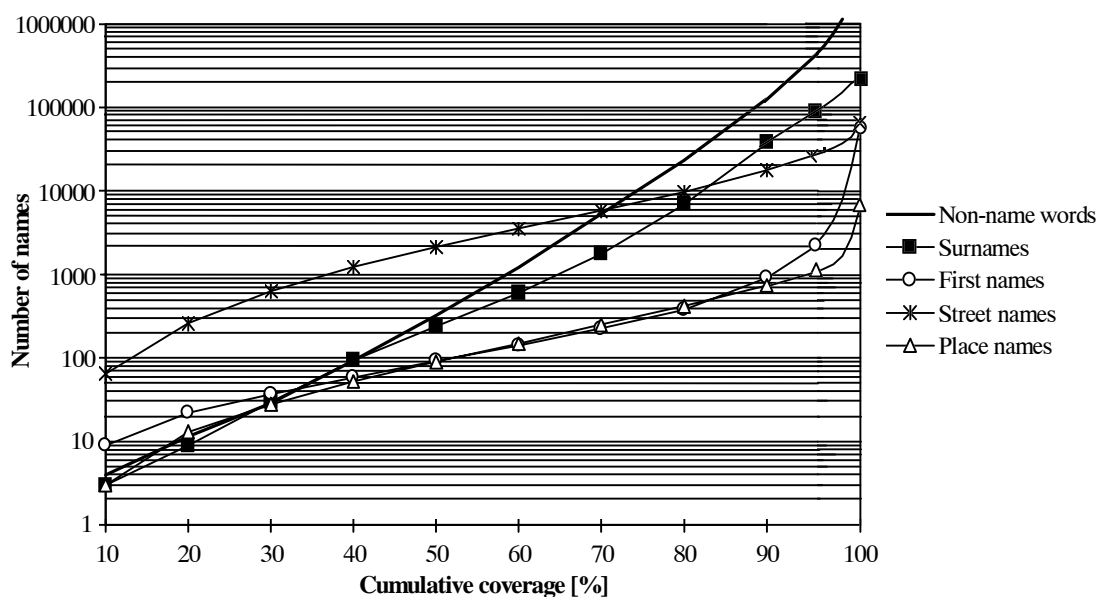


Figure 1. Cumulative coverage of names and non-name words in Sweden.

The next chapter will describe the structure of Swedish names. There are some terms that have to be defined to facilitate this:

What is a name?

The answer to this question depends on the requirements in a specific situation. Some definitions and their implications are listed in Table 5 below. You could either have a functional definition like the first definition, a pragmatic approach like the second, a descriptive definition like the third or a complex definition like the fourth.

Table 5. Different definitions of the term name.

Definition	Problem
1. Something used to refer to a person or place (Andersson, 1981)	Difference between name and appellative, e.g., mother is used in the same way as a name.
2. Something that identifies uniquely objects that belong to the same class (Yvon, 1993)	Potentially any string composed of figures and letters can become a proper name (telephone directory).
3. A name has no meaning, it does not give any information about the object it denotes (Mill, 1891; Gerritzen, 1993)	Sometimes names can tell something about the object, as in the case of technonyms and patronyms.
4. There are 3 criteria for names: i upper case initial letter ii unique and constant reference iii social convention (Wennstedt, 1995)	Some non-names are written with upper case initial letter. The social convention criterion is difficult to define.

What is the meaning of the same name?

The definition of the same name could be any one of these (WÄhlin, 1977):

- The same spelling** *Larsson* and *LARSSON* are the same names.
- The same pronunciation** *Hofberg* and *HÅvberg* are the same names.
This implies that it is known that *ëfi* used to be pronounced *ëví* after long vowel in Swedish and that *ëoi* and *ëÁí* can have the same pronunciation.
- The same origin** *Schmidt* is the German and *Smith* is the English version of the same name.

What kinds of names are there?

There are many different kinds of names. This is illustrated by the ten groups of names found by Wennstedt (1995): personal names, animal names, place names, company names, names of artefacts, mythological names, names of events, titles, product names and biological names. In the Onomastica project the following name categories were used.

- First name** Different terms for personal names are listed in Table 6.
- Surname** See Table 6 for descriptions and examples.
- Company name** The company names in the Swedish telephone book range from small snack bars to multi-national companies, to regional departments of trade unions. These names are disregarded in this thesis.
- Street name** Means address in the telephone book, which in most cases are streets or roads, but in some cases it is a very small village or even a harbour.
- Place name.** In the database these are equal to postal addresses.

Table 6. Listing of some terms used for personal names, (Lawson, 1984)

Term	Description, synonyms	Example
First name:	forename, given name, christian name or baptismal name	<i>John Q. Smith</i>
Middle name:	second name	<i>John Quincy Smith</i>
Surname:	last name, patronym, family name	<i>John Quincy Smith</i>
Nickname:	eke-name or sobriquet	<i>Red</i> for somebody with red hair
Affectionate:	familiar or adolescent	<i>Johnny Quincy Smith</i>
Hypocoristic:	short name or abbreviated name	<i>Ed</i> for Edward
Isonym:	individuals sharing the last name	<i>Sally Smith & Ed Smith</i>
Eponym:	a place named after a person	<i>Washington</i>
Technonym:	a parent derives his/her name from a child	<i>Um Daud</i> (mother of David)
Patronym:	a child derives names from father	<i>Larsson</i> (son of Lars)

What is a Swedish name?

Names have always moved with people across borders. Many names that are regarded as Swedish today have a foreign origin. Many German, French and Finnish names have been imported over the last 500 years. The German influence is reflected by the use of mute *ě* in names like *Wahlberg* and the French by, e.g., the use of *ěu* for *ě*. The ending of a name is often a good clue to the origin. The origins of some Swedish names are ambiguous due to the fact that many Swedish names have adopted a German ending like *-er*. In this study only names that have a German beginning as well as ending are regarded as German. All surnames have been automatically tagged with probable origin (for more details see chapter 6.1), tagging 63% of the names as Swedish, covering 92% of the subscribers. This figure is only approximate since the accuracy of the origin tagger is between 90% and 95%, according to a manual check.

4. The structure of Swedish names

Names have a different morphology and phonology compared to non-name words in Swedish, which makes them difficult to handle for ordinary letter-to-sound rules. There are many reasons why names have a different structure: many names are old, with old spellings. The spelling of Swedish words has been reformed a number of times, while many names have been left unchanged. Another reason is that names may have a foreign origin that influences the pronunciation to different degrees. Sometimes common names are spelled in unusual ways to make them more distinctive, for example using *ěí* instead of *ěsí* or inserting *ěhí*, e.g., *Cahrlzon* instead of *Carlson*, which is an alternative spelling of *Karlsson*.

This chapter will describe the different structures of Swedish names. Tables of statistics on non-name words and names are presented in the Appendix. The non-name word statistics are based on a corpus of 150 million words from newspapers, novels and other texts. The non-name words lexicon (NNW) has 107,379 lexical items, i.e., different word forms. In this study two different methods to compute statistics on names and non-name words have been used: non frequency weighted statistics, (NOFW), that gives information on which patterns are mostly used to construct words; and frequency weighted statistics, (FW), that gives information on which patterns used in the most common words. The statistics that are not frequency weighted are computed on the words in the lexicon. In the frequency weighted statistics each word has been weighted with the number of times it occurred in the corpus, for example the number of subscribers with the name *Eva* (49,978), or the number of times the word *ingjuta* occurred in the text corpus (107).

The phonetic transcriptions in the examples in this thesis use the International Phonetic Alphabet (IPA), with the exception that the stress mark is positioned before the stressed vowel instead of before the stressed syllable. This is done according to the standard used in the KTH text-to-speech system.

4.1. Surnames

The Swedish part of the Onomastica database contains 228,048 surnames, of which 46,859 are transcribed in band I and 82,983 in band II. The structures of the transcribed names have been analysed and the main results are presented in this chapter and in tables in the Appendix.

Swedish surnames are often multi-morphemic. This was for example shown in 1952 when an governmental surname committee studied 80,000 surnames. The names in this study included 11,500 different first morphs and 1,400 end-morphs. These figures can be compared to the 12,681 first morphs and 1,811 end-morphs found among the 109,001 transcribed Swedish surnames in the Onomastica database. The surnames have a uniform structure that can be divided into three main groups:

- I** Names that combine a male first name with the suffix *-son*.
The most common: *Johansson*, *Andersson*, *Nilsson*, *Karlsson* and *Ericsson*.
- II** Names that are compounds of two root-morphs, often nature related .
The most common: *Lindberg*, *Lindstr m*, *Lindgren*, *Lundberg* and *Bergstr m*.
- III** Others.
The most common: *Lundin*, *Bergman*, *Wallin*, *M ller* and *Blom*.

The structures of the Swedish surnames have a historical explanation (ModÈer, 1964). The ìson-namesî of category I were first used as patronyms. For example the son Sven of *Karl Andersson* was named *Sven Karlsson*. These began to be used as real family names during the 19th century. At this time *Johan*, *Anders*, *Nils*, *Karl* and *Erik* where among the most popular first names, which is the reason for the son-names with these to be the most common.

The names of category II have a uniform structure, with two compounded morphs that are both stressed on the first syllable. The morphs are often nature related nouns, but some adjectives occur as the first part. In the present corpus the first part is one of 6,952 morphs, whereas the second part is one of 1,421.

The bimorphemic names of category II have a number of sources. Some of them are noble names, from 1350 and later, which often describe the coat of arms. Since these often consist of compounded pictures or have multiple fields, the names became multi-morphemic, like *Bj^rnberg* which describes a coat of arms with a bear (*bj^rn*) on a mountain (*berg*). The morphs had to be heraldically and aesthetically acceptable. Typical initial noble name morphs are: *adler-* (Germ. ëagleí, 19 names in the lexicon have this morph); *gyllen-* (ëgoldení, 87); *norden-* (ënortherní, 121), and final: *-hj^olm* (ëhelmetí, 58); *-ski^ld* (ëshield, 159); *-stierna* (ëstarií, 58). Another group of noble names were those including the German *von* and its Swedish equivalent *af*. When these were introduced in the early 1700th century they established a connection between the family name and the family estate, e.g. *Boije af Genn^as*. In the 18th century they started to be used as a noble prefix to family names, without reference to the family estate, e.g. *af Enehjelm*.

Many names in category II are bourgeoisie names with two nature related root morphs. These are seldom joined by means of the ësí found in many types of noun compounds. For example *~kerberg* exists as a surname but not *~kersberg* (TegnÈer, 1882, Blomqvist 1993). The surnames of category II originated in the 17th century and were often derived from place names, e.g., *Lindberg* from the town *Lindesberg*. A more common pattern was a combination of the whole or part of a place name and with a nature related morph, e.g., *-gren* or *-quist*. An example of this are the names *Almgren*, *Almlind* and *Almquist*, where *Alm-* is derived from the village name *Almby*. Other examples are *Strindberg* from *Strinne* and *Wennergren* from the lake *V^onern*. Some of these two-morphemic bourgeoisie names were aimed at imitating noble names by adding a noble name morph or only the noble looking suffixes *-er* and *-en*, producing names such as *Engstr^mmer* from *Engstr^m* and *Langenski^ld* from *Lang*.

The most prominent reason for these names being bimorphemic is that they are generated from and in the same way as place names, which often are two-morphemic. The two-morphemic names normally have accent II, but in some parts of Sweden they have accent I, possibly due to German influence (ModÈer, 1964). The most common morphs in the names of category II are shown in Table a6 in the Appendix and in Table 7, together with their English translation.

Table 7. The five most common initial and final morphs in surnames of category II.

Initial morph	Final morph
BERG (ëmountainí)	BERG (ëmountainí)
AL (ëalderí)	MARK (ëgroundí)
DAL (ëvalleyí)	DAL (ëvalleyí)
STEN (ërockí)	G~RD (ëyardí)
LIND (ëlime-treeí)	FORS (ëstreamí)

Category III is the most varied name category. Some names are monomorphemic and come from place names, e.g., *Beck* from for example *Våstanbäck* or they are just plain nature related, e.g., *Berg*. Another large group of one-morphemic names are names of soldiers from the 17th century and later. These names were invented by the officers and were supposed to characterise the bearer. They were often plant- or animal-names, like *Dufva* (ëpigeoní) or *Lilja* (ëlilyí), but also adjectives like *Hård* (ëhardí) or *Glad* (ëhappyí). Sometimes the officers made up humorous names like *Tobak* (ëtobaccoí) or *Ruter* (ëdiamondí). Many of the stranger soldier names have disappeared over the years, but the names mentioned above are still in use.

Another group of names are constructed with certain name suffixes like *-ling* (1082) names in the lexicon have this ending), *-ner* (1842) and *-ler* (1296). These are often combined with place names, making names like *Meurling*, from the place *Mårhund*. Latin names endings, like *-elius* (993) *-enius* (840), *-erus* (84), *-inus* (18) and *-onius* (93) were popular among the clergy in the 16th and 17th centuries. These were either combined with the father's name, like *Svenonius*, or with a place name, like *Gavelius* from *Gävle* or from true translations of the whole or parts of Swedish surnames or place names, e.g., *Domerus* from the village *Husaby* (Lat. *domus*, Sw. *hus*). The names have their primary stress on the first syllable of the Latin endings, like *Dahlenius* [dal'e:nios]. These Latin names went out of fashion in the 18th century and the last part, *-(i)us*, was dropped, leaving names ending with *-ell* (2627), *-Ën* (3040), *-Ër* (545) and *-in* (3829). The new names kept the primary stress on the same, now last, syllable, like *DahlËn* [dal'e:n].

Other foreign endings of Swedish names are *-ander* (1333), which is derived from the Greek *ánēr* (ëmaní), and *-el* (1692), *-er* (3312), *-en* (253) and *-man* (2721) which are influenced by German names. The names with German influences have primary stress on the first syllable and accent I, for example *Bergman* [b'ærjman] and *Hedner* [h'e:dnər]. Names ending with *-er*, *-en* are ambiguous: they either have a German ending, *Leger* [l'e:gər] or they have dropped a Latin ending, *Legerius*->*Leger* [leg'e:r].

During this century many new names have been invented. The inventions of new names have been regulated by the authorities, starting with the introduction of a name law 1902 (Andersson, 1979). According to this law only names that are constructed, spelled and pronounced according to domestic language use are accepted as new surnames. Books with suggested new names following the name law were published. Two of these, written by Sahlgren in 1939 and 1940 consisted of unchanged place names, which have resulted in many names influenced by place names. This is illustrated by the fact that Tegnér d.y. (1882) found almost no surnames ending with the place name specific morphs *-rud*, *-ryd* and *-näs* while Modén (1964) refers to 776 *-rud*, 543 *-ryd* and 431 *-näs* names in 1954. In the 1994 telephone directory these numbers have increased to 995 *-rud*, 668 *-ryd* and 672 *-näs* names.

The constraints on new names have been relaxed in the last few decades, exemplified in the two latest books of name suggestions. In these books, from 1964 and 1979, computer software has generated new names from legal name morphs in Swedish. The morphs used were previously almost unused. Examples of initial morphs in these new names are: *Alt-* (189 Swedish and 53 foreign names are found in the telephone directory of 1994), *Kalm-* (59, 6), *Marm-* (60, 5). *Pors-* (79, 2). The requirement that the names must look Swedish has also been modified, giving names ending for example with *-ix*. There are 19 Swedish names ending with *-ix* in the database and 23 foreign names. Another change in the official policy has resulted in 40

son -namesî that include female names, e.g., *Gunillasson*, *Inezson* and *Hannason*, For a more detailed analysis of the structure of Swedish surnames see ModÈer (1964).

Only 2% of the names are of category I, but they have the greatest coverage, 42%. Most of the names, 61%, are of category II, and they cover 33% of all occurrences in the Swedish telephone book. The remaining 37% are of category III and they only cover 24%. Most Swedish surnames are disyllabic (53%) or trisyllabic (40%), as can be seen in Figure a3 in the Appendix. Compounds of surnames, such as *Sundgren-Svensson*, have been split in the database, which explains why there are no surnames with more than five syllables. Almost all surnames with accent II have their main stress on the first syllable (99.6%), as most Swedish words of that accent. The position of the primary stress varies more in surnames with accent I, where 56% have primary stress on the first syllable and 42% on the second.

Table 8. Some statistics on patterns in surnames.

Pattern	In surnames	Not in NNW
Diphones	1,265	4
Triphones	11,182	2,407
CVC-patterns	406	13
Stress patterns	38	0

There are six diphones in the surnames that were not found in the non-name words lexicon, as can be seen in Table 8. These diphones and a surname example are shown in Table 9. The first three are probably of German origin. The Swedish ones are - TENSE high front vowels followed by a retroflex ädí or ëní.

Table 9. Names that contained diphones that were not found in common word.

Surname	Transcription	Diphone
H _u bsch	h'ʏbf	bʂ
Hamsch	h'amf	mʂ
Jirdell	jɪd'el	ɪd
MyrdÈn	mɪd'e:n	ɪd

The 10 most common CVC-patterns for surnames have a cumulative coverage of 47.6% (NOFW) and 56.7% (FW), as can be seen in Figure a7 in the Appendix. The correlation of the rank of the 10 most common CVC-patterns in surnames with the corresponding ranks for common words is shown in Figure a9. The CVC-patterns for surnames have approximately the same rank in common words (NOFW). In the frequency weighted top-10 list there are some CVC-patterns that are rarer in non-name words. For example CVVCCVC rank 4 in surnames but 207 in non-name word. Example of names and words with this pattern are:

surnames:

Johanson Leander Boestad
 j'u:ansøn le:'andər b'u:ə-stɑ:d

non-name words:

fiender teatrar diagnos
 f'i:endər te'a:trar dragn'o:s

There are 13 CVC-patterns that occur in surnames, but not in any word in our lexicon of common words. These are listed in Table 10. Two characteristic features are found in these patterns. The first is the occurrence of multiple consonant clusters. These are caused by the compounding of name morphs with consonant clusters, like

Björk-strander. The second feature is the occurrence of vowel sequences, for example in *Eurenius*. Many of these are Swedish names with the Latin ending *-ius* or names of German or Norwegian origin containing *-oe* and *-au*.

Table 10. Surnames with transcription that contains cvc-patterns that do not occur in the lexicon of Swedish common words.

Surname	Transcription	CVC-pattern
Björkstrander	bj"œrk-str, andər	CCVCCCCCVCCVC
Strandbro	str"and-br,u:	CCCVCCCCV
Strandkvist	str"and-kv,ist	CCCVCCCCVCC
Strömsmoen	str"øms-m,u:ən	CCCVCCCVVC
Strandelius	strand'e:liəs	CCCVCCVCCVVC
Eurenius	eør'e:niəs	VVCVCCVVC
Stranius	str"ɑ:niəs	CCCVCCVVC
Strönder	strø:'andər	CCCVCCVVC
Hyenstrand	h"y:ən-str, and	CVVCCCCVCC
Graufelds	gr'aøfelds	CCVVCVCCC
Braunerhjelm	br"æonər-j, elm	CCVVCVCCVCC
Haugbergsmyr	h"æog-bærjs-m,y:r	CVVCCVCCCCVC
Underdalshaugen	"øndər-dɑ:ls-h, æogən	VCCVCCVCCCCVCCVC

There are only 38 different stress patterns in Swedish surnames, compared to 500 for common words. The seven most common stress patterns (NOFW) cover 88.6% of all names, and they can be used to summarise the different types of surnames described in this chapter:

1. "V-V *Lindberg* nature related two-morphemic bourgeoisie name (II)
2. "VV-V *Nordenskiöld* two-morphemic noble name (II)
3. 'VV *Bergman* name with German influenced ending *-man* (III)
4. VV *Lundin* name with Latin ending *-inus*, later reduced to *-in* (III)
5. "VV *Duva* soldier name (III)
6. 'V *Blom* nature related one-morphemic name (III)
7. "VVV *Andersson* son-name (I)

4.2. First names

There are 60,850 first names in the database, of which 35,800 occur only once. The 10,479 most common first names were transcribed and analysed. Table 11 shows the five most common male and female first names in the database and in a Swedish first name book (Allén, 1995). There are two main reasons for the difference between the two: the first name book counts all given names of a person, while the telephone book often lists one of them. Secondly the first name book includes almost all 8.5 million Swedes, of all ages, while the telephone directory only includes 4.5 million subscribers, with a different age and sex distribution. There are for example more male subscribers, because it is often the husband in the family that is listed in the telephone book.

Table 11. The most common male and female first names in the Swedish telephone directory and in the first name book.

Telephone book		First name book	
Male	Female	Male	Female
Lars	Eva	Erik	Maria
Anders	Karin	Karl	Anna
Lennart	Kerstin	Lars	Margareta
Bengt	Lena	Anders	Eva
Gunnar	Ingrid	Per	Elisabeth

The names were tagged with sex in the morphological analyser lexicon, giving 2142 of the first names in the lexicon a male tag and 2,475 a female tag. These tagged names cover 97,4% of the occurrences in the Swedish telephone directory. The stress-patterns of male and female first names are shown in Table 12. In the first column the stress pattern of the name is described by two numbers, x:y, where x is the number of syllables and y is the syllable that carries the main stress.

Table 12. The stress pattern of male and female Swedish first names.

Stress pattern	Male names			Female names		
	NOFW	FW	Example	NOFW	FW	Example
1:1	7 %	31 %	Bo	3 %	5 %	Ann
2:1	35 %	58 %	Arne	38 %	51 %	Eva
2:2	1 %	0 %	RenÉ	6 %	8 %	Mari
3:1	4 %	3 %	Mikael	20 %	12 %	Annika
3:2	50 %	8 %	Johannes	9 %	15 %	Agneta
3:3	0.4 %	0 %	Severin	5 %	1 %	Josefin
4:1	0 %	0 %	Veli-matti	11 %	2 %	M ata -stina
4:2	0 %	0 %	Karl-mikael	2 %	2 %	Elisabeth
4:3	0 %	0 %	Alexander	2 %	3 %	Margareta

Most men and women have a name consisting of two syllables with the primary stress on the first syllable. Most male names are trisyllabic with primary stress on the second syllable, of which most are double names with primary stress on the second name. The method of creating new compound names from existing ones is a very productive one. 50% of all male names in the dictionary are compound names, but these only cover 7% of the subscribers. The large number of double names explains why most male names have the primary stress on the second syllable. The most used male names have their primary stress on the first syllable. Most female names have the primary stress on the first syllable.

Double names are common among both the male first names (50%) and the female first names (30%). These have a uniform pronunciation structure, but the structure of male and female double names differs, see Figure 2. Male double names are often trisyllabic with accent I and stress on the second syllable. The female double names vary more, but often have three or four syllables with accent II and primary stress on the first syllable, following the normal pattern of Swedish compounds. First names are tagged with sex in the morph lexicon to handle this difference in pronunciation.

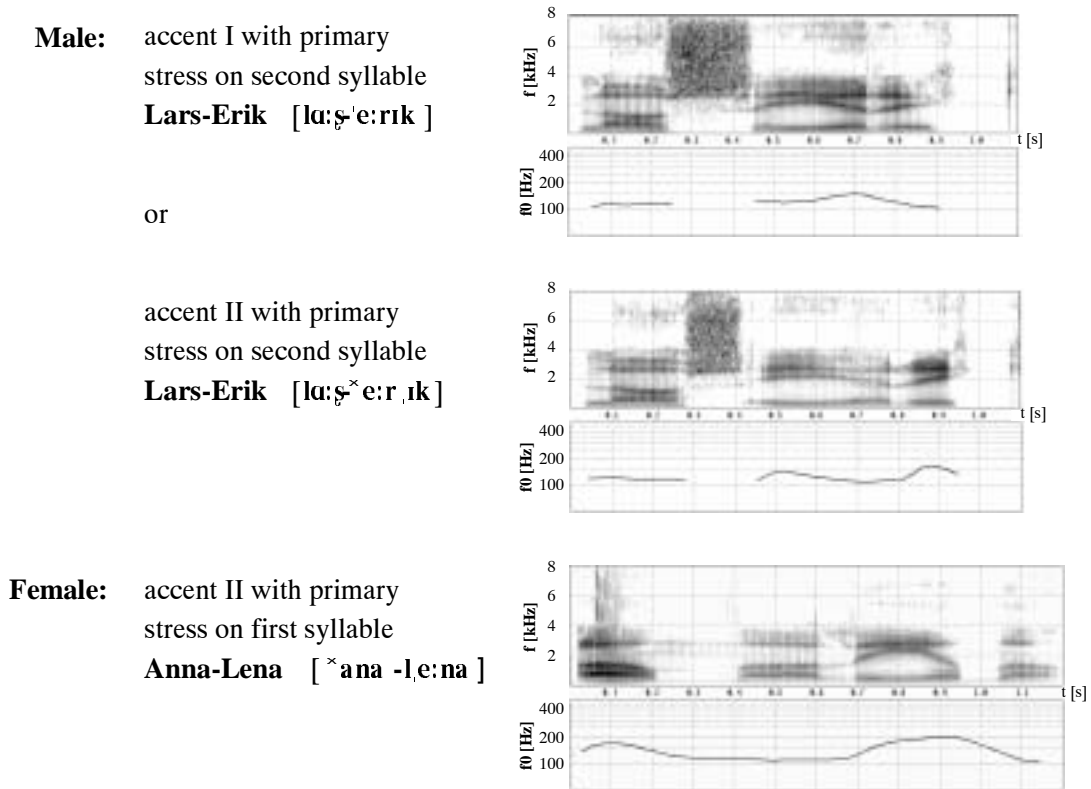


Figure 2. The stress-pattern of male and female compound first names

The difference in pronunciation of male and female double names has been explained by NorÈen (1907). The first part in male double names have lost their stress in analogy with the pattern of first name + family name, for example the name *Karl Andersson* is pronounced [kɑ:|ˈandəɕɔn], and in the same way the double name *Karl-Anders* is pronounced [kɑ:|ˈandəɕ]. Men were often called by their first name followed by their family name, while women usually were called only by their first name. This is why the analogy has not been used for female double names.

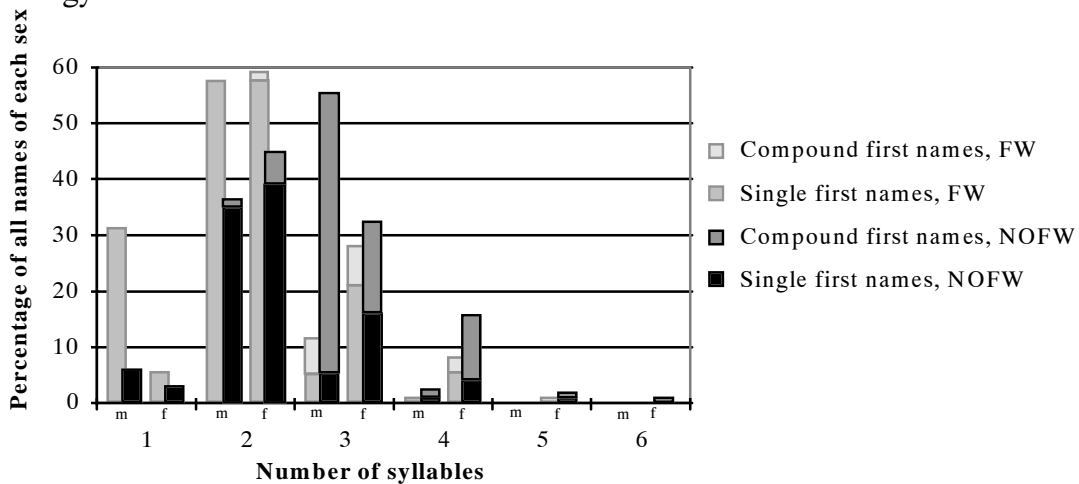


Figure 3. The distribution of number of syllables for male and female first names

Figure 3 shows the number of syllables for male and female single and double names. Most names have two or three syllables. Only 7% of the male names are single-syllable names, but they have a coverage of 31%. Most names that have four syllables or more are female. The male first names contained an average number of 1,8 syllables and the

female 2,4 syllables. The difference is 0,6 syllable which is slightly less than reported in earlier studies (Otterbjörk, 1979). They found that female names were almost one syllable longer than male names.

Earlier in this chapter it was shown that the most common stress pattern for both male and female names was a disyllabic name with the primary stress on the first syllable. Those with names of other lengths often get affectionate forms of their names, that almost always are two-syllabic, with primary stress on the first syllable and accent II. Elert (1964) proposed that this pattern has an expressive function in Swedish, used in pet names, e.g., *Maggan* for *Margareta*, in nursery words, e.g., *fosse* for *fot* (ēfootí) and in slang, e.g., *fralla* for *franskbröd* (ēFrench breadí). The endings *-an* and *-is* are often used to generate these words (Ståhle, 1979). In Stockholm many place names have a slang form, e.g., *Rålambshov* → *Rålis*; *Tekniska Högskolan* → *Teknis*; *Stallmästaregården* → *Stallis*. The ending *-is* is also used in non name words as *dagis* (ēday nurseryí) and *kändis* (ēcelebrityí).

The affectionate forms of the names are mostly ellipses generated by extracting the beginning of the word, often as far as the second vowel. This vowel and the rest of the names are replaced by an expressive gemination and an affectionate ending (Eliasson, 1979). Some examples are:

<i>Sigurd</i>	->	<i>Sig+ge</i>	->	<i>Sigge</i>
<i>Joakim</i>	->	<i>Jo(a)k+ke</i>	->	<i>Jocke</i>
<i>Frans</i>	->	<i>Fra(n)s-se</i>	->	<i>Frasse</i>
<i>Malin</i>	->	<i>Mal+la</i>	->	<i>Malla</i>
<i>Viktoria</i>	->	<i>Vik+kan</i>	->	<i>Vickan</i>
<i>Margareta</i>	->	<i>Ma(r)g+gan</i>	->	<i>Maggan</i>

The most common endings in these affectionate names are shown in Figure 4. The most common male ending is *-e*, and most common female is *-an*.

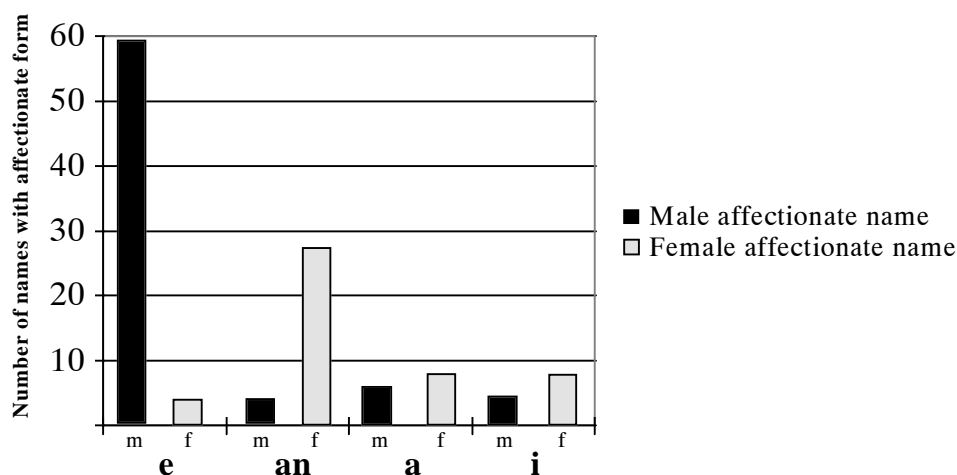


Figure 4. The distribution of the endings in affected forms of names.

The 100 most common names of length 1, 2, 3 and 4 were selected to study the use of affectionate forms. These 400 names have a coverage of almost 60%. The author of this thesis tried to find affectionate forms for these names. The distribution of affectionate forms for the names of different lengths and stress patterns can be seen in Figure 5. The figures to the left of the arrows show the number of names that got the affectionate form and the figures to the right give the coverage of these names.

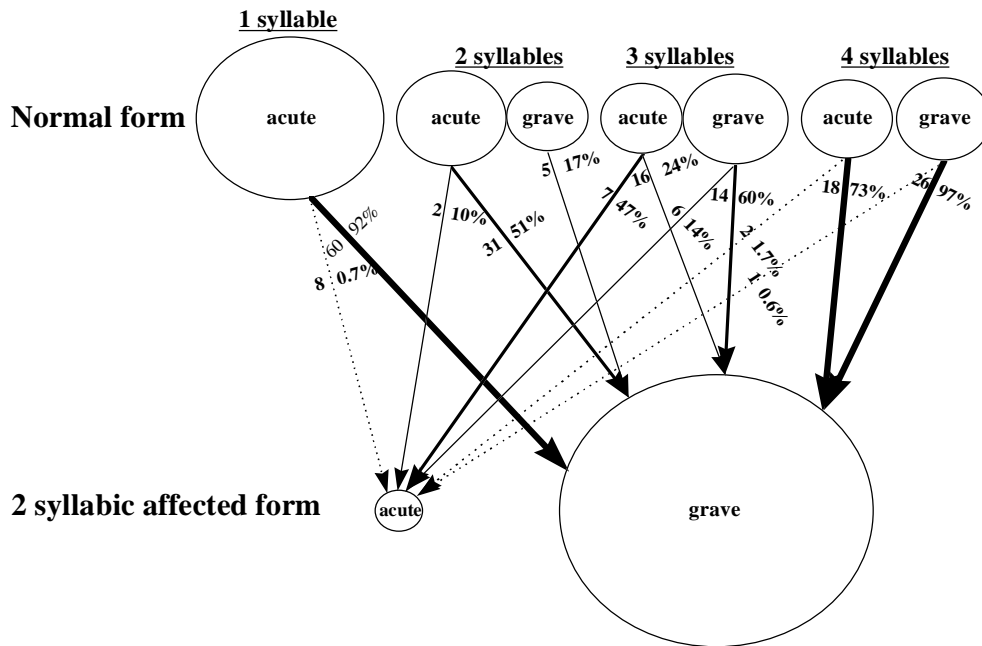


Figure 5. The number of names of each length and stress-pattern type that have a common affectionate form. The circle-sizes are proportional to the number of names in that group, and the line-weights indicate the coverage of the names.

Notice that almost all names with one or four syllables have a disyllabic affectionate form. Many trisyllabic names have an affected form, while most disyllabic names do not. However, the disyllabic names that got an affectionate form were accent I names, with an affectionate form of accent II, e.g., *Fredde* for *Fredrik*. The ones that were unchanged were already in the preferable form.

The first names have been analysed and compared to the non-name words. The main results are shown in Table 13. For more detailed statistics see the Appendix.

Table 13. Some statistics on patterns in first names.

Pattern	In first names	Not in NNW
Stress patterns	33	0
CVC-patterns	195	6
Diphones	754	5
Triphones	3,074	495

All stress patterns found in first names can also be found in non-name words. There are six CVC-patterns that only occur in first names:

Claudia	Georgia	Ann-Sofia	Alexia	Ia	Eugenia
k'læðɪɑ	j'e'ɔrgɪɑ	˜an-suf'i:a	al'eksɪɑ	˜i:a	eøʃ'e:nɪɑ
ccvvcv	cvvccv	vccvcv	vcvccv	vv	vvcvcv

Notice that all these CVC-patterns contain vowel sequences. Most of them are either of foreign origin, e.g., *Alexia*, or compounds, e.g., *Ann-Sofia*. This explains why these CVC-pattern are not found among the 107,379 most common Swedish non-name words. The five diphones that were only used in first names were [æɪ, æd, æg æt, æŋ]. All of these occur in English names like *Janice*, *Gladys*, *Maggie*, *Kathleen* and *Franklin*. The Swedish [æ] only occurs before retroflex consonants in standard Swedish.

4.3. Place names

Most place names are two-morphemic (79%) or one-morphemic (16%), a few are three-morphemic (5%). The two-morphemic names are similar to surnames, mainly because many surnames are derived from place names. The first morph in the two-morphemic names is one of 2542 nouns or adjectives and the second morph is one of merely 542 nouns. This is similar to the surname-pattern, but the morphs differ. The 50 most common morphs in bimorphemic place names and surnames are shown in Figure a6 in the Appendix. Figure 6 shows the overlap of morphs between place names and surnames, as well as the overlap between morphs in first part and the second part of each name type.

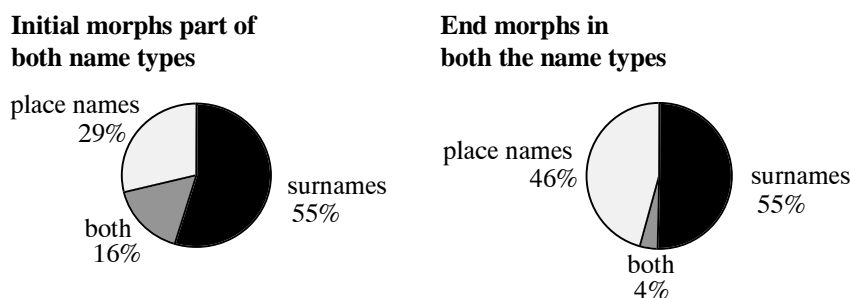


Figure 6. The overlap of morphs in two-morphemic place names and surnames. The surnames have 50% more morphs than the place names.

Seemingly the first morph makes the name unique while the second morph is used to identify the name as a place name or a surname. The initial morphs of the surnames have quite a large overlap with morphs in all other positions, see Figure 7. The overlap between the end-morph in surnames and the end-morph in place name is much smaller than the overlap between the two initial morph groups, which have the largest overlap. This supports the idea that it is the second part that characterises the names. There are some other differences between them. Table a3 in the Appendix lists the 50 most common morphs of the four types described above. As can be seen in the table, the spelling differs: surnames are spelled with *öhí*, *ëqí* and *ëzí*, while place names mostly keep the spelling of the original words.

The stress patterns of the bimorphemic names of these types also differ. Bimorphemic surnames almost always have the structure of common compounds of two root-morphemes, e.g., *Bergstr^ˆm* [b^ˆæ:ɾj-str^ˆɸm]. Many place names that consist of two root-morphs have changed their pronunciation to resemble one-morphemic words with a derivational suffix, e.g., *Malm^ˆ* [m^ˆalm^ˆɸ]. The same name can be either a place name or a surname, but with different stress patterns, e.g., *Gr^ˆndal* will be pronounced [gr^ˆɸ:n-d^ˆɑ:l] when it is a surname; and [gr^ˆɸnd^ˆɑ:l] when it is a place name. The different stress patterns of two-morphemic place names are shown in Table 14.

Table 14. Some statistics of patterns in place names.

Place names	Stress patter
2,875	accent II with morph boundary (^ˆ M#M)
807	accent II without morph boundary (^ˆ MM)
582	accent I with primary stress on first morph (^ˆ MM)
452	accent I with primary stress on second morph (M ^ˆ M)
58	accent II with primary stress on second morph (M ^ˆ M)

Statistics on some patterns are displayed in Table 15. Only one diphone, [uŋ], occurs exclusively in place names, as in the Lappish *Vivungi*. No stress patterns are unique for place names, and only 20 unique CVC-pattern, were detected. Of these, four come from Lappish names, e.g., *Kieksi%oisvaara* [k'ɛksɪɛjsvɑ:ra] CVCCCVCCCV, eight are two-morphemic, e.g., *flands-bro* [ɛ:landsbr'u:] VCVCCCV and eight are three-morphemic, e.g., *Bergs-viks-holme* [bærjsvi:ksh'ɔlmə] CVCCCVCCCV. The 10 most common CVC-structures, shown in Figure a16 in the Appendix, are quite frequent in non-name words as well. However, the fourth most common pattern, CCVCCVCC, has rank 100 in non-name word CVC-patterns. This pattern occurs in *Stockholm* [st'ɔkhɔlm] and *TrÂngsund* [tr'ɔŋ-sʊnd]. Examples of common words with this pattern are *tveksamt*, *praktisk* and *trÂskeln*. Most of the missing patterns would be found in a larger lexicon that includes more compounds and if co-articulation rules are applied across morph boundaries.

Table 15. Some statistics on patterns in place names.

Pattern	In place names	Not in NNW
Diphones	1,086	1
Triphones	6,153	905
Stress patterns	57	0
CVC-patterns	446	20

4.4. Street names

The database contains 64,621 street names, of which 39,822 were transcribed and analysed. The most common street names are: *Storgatan*, *Ringv%ogen*, *Kungsgatan*, *Nygatan* and *Skolgatan*. Street names generally consist of one or two common words followed by *-gatan* (ëstreetí) or *-v%ogen* (ëroadí), for example *Gr^n-dals-v%ogen* (ëGreen-valley-roadí). The 39,822 street names were processed by Twol, giving parts of speech tags on the included morphs in 29,518 of the names. Table 16 shows the most common patterns for these street names. The 10 most common patterns in the first part of the street names cover 98% of all street names. These patterns are followed by ëstreetí, ëroadí or another of 400 end-morphs. The 10 most common end-morphs in Table 17 cover 95% of all street names.

Table 16. The most common patterns in the first part of street names.

Pattern before end-morph	Coverage (%)
Noun#Noun	46.0
Name	27.0
Noun	7.0
Adjective#Noun	4.0
Adjective Noun	3.0
Adjective	1.0
Adjective Noun#Noun	0.6
Adjective Name	0.5
Verb#Noun	0.3
Noun#Verb	0.2

Table 17. The most common endings of street names.

End-morph	Coverage (%)
v%ogen (ëthe roadí)	46.5
gatan (ëthe streetí)	27.0
v%og (ëroadí)	7.8
gr%ond (ëalleyí)	4.9
stigen (ëthe pathí)	3.6
backen (ëthe hillí)	1.8
gata (ëstreetí)	1.6
gÂngen (ëthe pathwayí)	0.7
plan (ëplaceí)	0.7
torget (ëthe squareí)	0.5

In the two-morphemic street names the first part mostly consists of a name, a noun or an adjective. The three-morphemic street names have a similar pattern: the first part is usually a noun or adjective, whereas the second part almost always is a noun in genitive form. Five of the most common initial morphs are shown in Tables 18 and 19.

Table 19. Common initial morphs in two-morphemic street names.

Table 18. Some common morphs in three-morphemic street names.

First morph			First morph		Second morph
Name	Noun	Adjective	Noun	Adjective	Noun
Balder	skog (ēforestí)	v%oster (ēwestí)	sten (ēstoneí)	ny (ēnewí)	bergs (imountainísí)
Sleipner	kvarn (ēmillí)	lÂng (ēlongí)	sĵ (ēlakeí)	stor (ēbigí)	gÂrds (iyardísí)
TegnĒr	sĵ (ēlakeí)	ny (ēnewí)	berg (ēmountainí)	hġ (ēhighí)	torps (icroftísí)
Idun	strand (ēbeachí)	stor (ēbigí)	sand (ēsandí)	norr (ēnorthí)	%ongs (imeadowísí)
Frġding	sol (ēsuní)	sġder (ēsouthí)	sol (ēsuní)	lÂng (ēlongí)	Âs (iridgeísí)

As can be seen in Table 20 there are only 16 unique diphones, most of which are generated at compound boundaries. Most of these occur at morph boundaries, which explains why they did not occur in the common word lexicon. When a lexicon of 1.8 million common words is used, all of them are found. The 10 stress-patterns that only exist in street names have two main origins. Firstly, street names are constructed by compounding words that are not usually compounded, e.g., *Holl%ondare#hus#v%oge*. Secondly multiple words are used as street names, e.g., *V%ostra J%ornv%ogsesplanaden*. This explains the large number of CVC-patterns that are unique for street names.

Table 20. Some statistics on patterns in street names.

Pattern	In street names	Not in CW
Diphones	1,299	16
Triphones	13,978	2,998
Stress patterns	140	10
CVC-patterns	2,251	939

Certain rules have to be included in an automatic transcription system for street names. Since street names often consist of multiple words, they are often abbreviated. Names beginning with adjectives for the points of the compass, *norra*, *sġdra*, *v%ostra* and *ġstra* are abbreviated *ġn.í*, *ġs.í*, *ġv.í* and *ġ.í*. Other adjectives and titles like *ġdoctorí* are also abbreviated. The endings *-gatan* and *-v%ogen* are mostly written *-g.* and *-v.* The rules for expanding these abbreviations are shown in Table 21.

Table 21. The use of the street name-endings *-g.* and *-v.*

The context of the endings <i>-g.</i> and <i>-v.</i>	Example of expanded street names
Most names ending with <i>-g.</i> and <i>-v.</i> are in definite form	<i>Storg.</i> - <i>Storgatan</i> <i>Ringv.</i> - <i>Ringv%ogen</i>
If the last word is <i>skolg.</i> , <i>byg.</i> , <i>kyrkog.</i> or similar it is in indefinite form	<i>Maria skolg.</i> - <i>Maria Skolgata</i> <i>SpÂnga kyrkv.</i> - <i>SpÂnga kyrkv%og</i>
If the ending is preceded by an adjective and space it is in definite form	<i>Lugna g.</i> - <i>Lugna gatan</i> <i>Sġdra v.</i> - <i>Sġdra v%ogen</i>
If the ending is preceded by a name and space it is in indefinite form	<i>Renstiernas g.</i> - <i>Renstiernas gata</i> <i>Drottning Kristinas v.</i> - <i>Drottning Kristinas v%og</i>

Other problematic abbreviations are initials in names. There is no way to expand the name *J A Perssonsg.* if you do not know who the person behind the name is. Roman numbers are used in names of kings, for instance *Christian IV v.* for *Christian den fjördes vög.*

Some etymological features change the pronunciation of a name. There is for example a difference in pronunciation of bimorphemic surnames and place names, as was previously shown. Bimorphemic surnames always have primary stress on the first morph. The place names differ more, for example names ending with *-berg*, *-lund* and *-dal* will mostly get primary stress on the last morph, for example *Katrinelund*, *Karlberg* and *Gr̂ndal*. If a street name consists of a place name followed by *vögen* and it begins with an adjective, like *norra* (ēnorthern), this could either relate to the street or to the place. The adjective will be de-accentuated if it relates to the place name. The street ÷vre *Gr̂ndalsvögen* can get three different readings, depending on the etymology of the name, see Table 22. The fourth possible reading, of a person called ÷vre *Gr̂ndal* is regarded as incorrect.

Table 22. Alternative pronunciations of ÷vre *Gr̂ndalsvögen* and their etymology.

Etymology	Transcription
there is a place called ÷vre <i>Gr̂ndal</i>	[ø:vrə-grø:ndʰɑ:ls-v̥ɛ:gøn]
there is also another street <i>Nedre Gr̂ndalsgatan</i> , that is generated from a place <i>Gr̂ndal</i>	[ʰø:vrə grø:ndʰɑ:ls-v̥ɛ:gøn]
there is also another street <i>Nedre Gr̂ndalsgatan</i> , that is generated from a person with surname <i>Gr̂ndal</i>	[ø:vrə grʰø:n-dɑ:ls-v̥ɛ:gøn]

5. Grapheme-to-phoneme conversion

The correct conversion from a string of letters to its pronunciation is essential in a text-to-speech system. The simplest approach would be to have a pronunciation dictionary, but this would have to be extremely big in a general system, especially if names are included. The use of a domain dependent dictionary is often used. In languages such as Swedish where compounds are very common, some grapheme-to-phoneme conversion technique will be needed as well, for example context dependent rules or a morphological decomposition. There are a number of techniques that have been used for this task, of which the most used is the rule-based method. However, rule-based systems have often proven to be insufficient when dealing with names. The different approaches used for transcription of names in the Onomastica project and elsewhere will be presented in this chapter.

5.1. Context dependent rules

The context dependent rules used in the Swedish part of the Onomastica project are inspired by the formalism of generative phonology. A dictionary of aligned grapheme-phoneme strings is often used to derive the transcription rules. Any string of N graphemes aligned with a string of M phonemes can be described by at the most N context dependent rules. This would be the case where you simply have one rule for each word generating the transcriptions. It is quite easy to extract rules in this way, but the problem is to minimise the number of rules, for example by unifying rules that share the same context. The rule-based system at KTH (Carlson and Granström, 1976) uses context sensitive transformation rules in a notation that is similar to the one used in generative phonology. The structure of these is exemplified by the hypothetical rules in Figure 7.

1: A \rightarrow A / B _ R	Ex. BAR \rightarrow BA:R
	BARRA \rightarrow BA:RRA
	BARSK \rightarrow BA:RSK
2: A: \rightarrow A / _ <CONS>(2,3)	Ex. BA:R \rightarrow BA:R
	BA:RRA \rightarrow BARRA
	BA:RSK \rightarrow BARSK

Figure 7. The RULSYS notation for context dependent rules.

Rule number 1 replaces all occurrences of A after B and before R with A:. Features like vowel, nasal or stress can also be used in the rules. Rule 2 uses the feature consonant, <CONS>, where it replaces all occurrences of A: before two or three consonants with A. The rules are applied in the order they are written. This makes the order of the rules important, as can be seen in Figure 8. In this example the same two rules will result in different output because rule number 2 changes the context used in rule number 1.

1: A \rightarrow C / C _ C	Ex. CACA \rightarrow CCCA
2: C \rightarrow A / _ A	CCCA \rightarrow CCAA
1: C \rightarrow A / _ A	Ex. CACA \rightarrow AAAA
2: A \rightarrow C / C _ C	

Figure 8. The result of applying the same two rules in different order.

In the Swedish rule system obvious stress marking and consonant changes are done in the beginning, while more sophisticated rules that depend on other rules appear later. Morph decomposition is done during this process, since it is important for the pronunciation of Swedish word, for example when determining of the accent of the word.

5.2. Symbolic learning

A system that uses the symbolic learning technique learns to do grapheme-to-phoneme conversion by training on a large corpus of aligned pairs of grapheme strings and phoneme strings. The aligned corpus must be manually corrected. The training results in a stochastic decision tree. Decision trees are like ordered lists of context dependent rules. However, for languages like English this approach will probably produce transcriptions with a phoneme accuracy of at most 90-95%, which corresponds to a word accuracy of 55-60%. Symbolic learning is used in the Danish part of the Onomastica project. Their table-lookup approach is called SELEGRAPH. (Anderson and Dalsgaard, 1995). An overview of the system is shown in Figure 9.

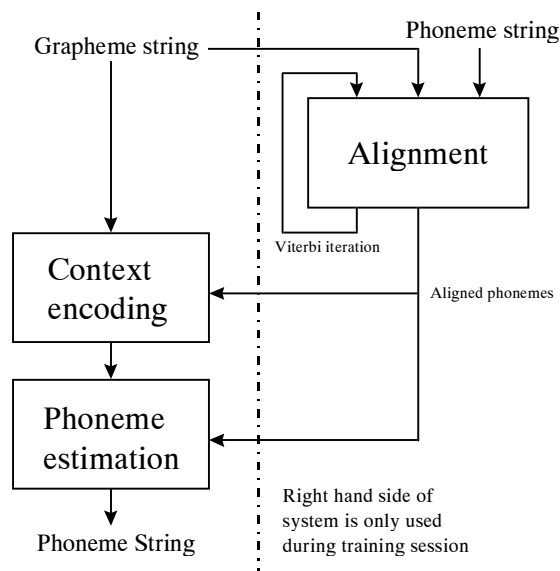


Figure 9. Architecture of the SELEGRAPH system (Anderson and Dalsgaard, 1995).

The system is trained in three steps:

- First the grapheme strings and phoneme strings are aligned using an iterative Viterbi alignment algorithm.
- Then the mutual information from the context is computed, enabling determination of the size of the input window and in which order the context is to be considered.
- Finally the stochastic decision tree is built by examining a dictionary of aligned pairs of grapheme and phoneme strings. The tree consists of leaves and branches. The leaves contain statistics of possible phonemes for each grapheme in a given context. Each leaf has a pointer to the following leaf.

Default mappings are used for unseen words with grapheme strings that were not in the dictionary and for ambiguous conversions.

When the system is used for transcription the tree is traversed from the root down the branches in the order given by the mutual information, until the desired grapheme string is obtained. On the way down each leaf outputs phonemes according to the associated statistics.

This approach has problems with generalisation, since only default mappings are used for unseen grapheme strings. Another drawback is that it could output transcriptions that are not legal, for example transcriptions with more than one primary stress. The word accuracy on surnames for this system ranges from 66% for Danish to 95% for Italian.

5.3. Artificial Neural networks

An Artificial Neural Network (ANN) is a computational model, that is influenced by the biological neural networks of the brain. An ANN consists of a number of nodes in layers that are connected to nodes in other layers. The weights of the connections are trained by presenting input data at the input layer and then adjusting them until the activities in the output layer are close enough to some predefined wanted output activities. ANNs have mostly been used for speech recognition, but there are some examples of grapheme-to-phoneme conversion ANNs. The ANN learns to transcribe words by being trained on a dictionary. The network processes the letters one by one and outputs a set of activation levels. These are mapped to the closest bit string that represents a legal phoneme/stress pair. The most famous system that uses this technique for letter-to-sound conversion is NETtalk (Sejnowski and Rosenberg, 1987). The Portuguese Onomastica system (Viana et al., 1995) is inspired by this system, and it uses the conventional multi-layered feed-forward neural network. Their system produces transcriptions with word error rates of 7.3% for a test-set of common words and 12.4% for names.

A problem with the neural network approach is that it does not cope very well with irregular words, which names often are. Another problem is that ANNs produce transcriptions letter-by-letter, where the best result so far for languages with complex spelling like English is a phoneme accuracy of around 90-95% (Ainsworth and Pell, 1989), giving a word accuracy of around 50-60% (Coker et al, 1993).

5.4. Markov Models

Hidden Markov Models (HMM) are often used in speech recognition, but they can be modified to do grapheme-to-phoneme conversion as well (Parfitt and Sharman, 1991; Rentzepopoulos and Kokkinakis, 1991; Riley, 1991). The HMM model is based on the statistical properties of the task. A Markov process is a chain of states, in this case the phonemes. Transitions between the states have probabilities, for example the probability a_{ij} to move from state i to state j . At each state the letters are emitted as observable outputs, while the states themselves are hidden. The task of the system is to determine which hidden state sequence (the transcription) that was most probably traversed to generate a given output sequence (the orthography). This will produce the most probable transcription of the given orthographic string. An example with three names is shown in Figure 10. In this example the probability that grapheme ëYí is emitted in state six is b_y while the probability for emitting ëIí is b_r . State four is a null-phoneme.

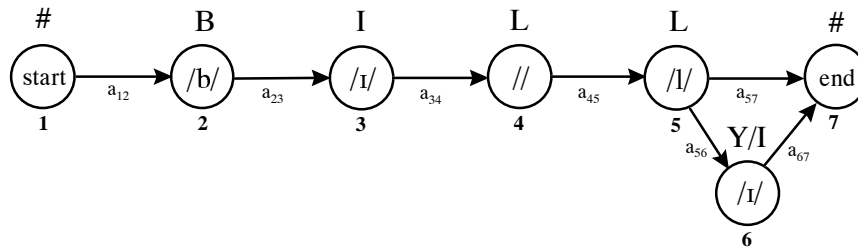


Figure 10. A Markov model of the names Bill, Billy and Billie.

The system is built in four steps :

- First the appropriate states and outputs have to be defined.
- Then the initial transition and output probabilities are estimated, using a corpus of aligned grapheme and phoneme strings.
- The most probable state sequence to produce a given output is obtained using the Viterbi algorithm. The performance of the model can then be computed by comparing the state sequence with the correct transcription.
- Finally the initial parameter estimations are improved by using the Forward-Backward algorithm in a number of iterations.

When the system is used for transcription the Viterbi algorithm is used to get the most probable phoneme sequence. If the system is to be used for phoneme-to-grapheme conversion the only thing that has to be done is to swap the grapheme string with the phoneme string in each state.

The HMM approach has some weaknesses: One is that it cannot use longer range context. These are quite common in names, where the ending correlates with the origin of a name, and therefore influences the transcription of the whole name. Another drawback is that HMMs tend to over-train on the training-data, giving insufficient generalisation. Finally appropriate estimations of the transition and output probabilities must be decided for all unseen sequences.

5.5. Analogy

The analogy approach uses a lexicon with transcribed words to create transcriptions for unknown words. The strength of analogy has been shown by Coker et al. (1993) and Sullivan and Damper (1991). This approach has been used in the Italian Onomastica system (Pirelli and Federici, 1995). Their system is described in Figure 11. First the system must find out which grapheme strings to match with which phonemic strings. These matches can be anywhere in the words.

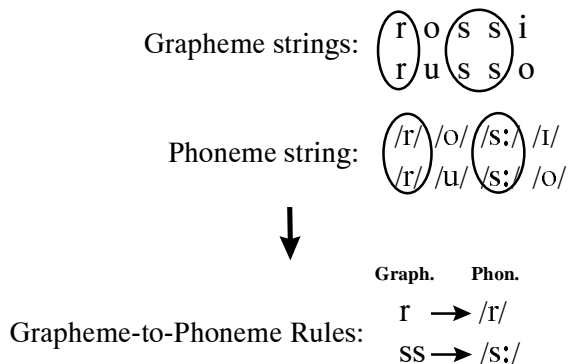


Figure 11. Learning two mappings by example (Pirelli and Federici, 1995).

The systems always favour longer grapheme strings over shorter ones. This approach is possible when dealing with a language like Italian where the mapping from grapheme to phoneme is quite consistent. In languages like English and Swedish, however, where the grapheme-to-phoneme correspondences are far less regular, it is necessary to have a large dictionary and use context. Otherwise the system could generate transcriptions like the classical example by G. B. Shaw shown in Figure 12.

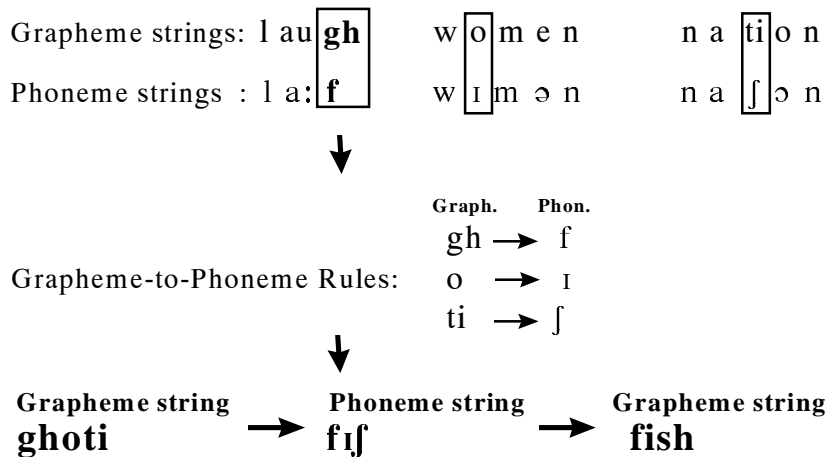


Figure 12. To pronounce *ęhotiü* in the same way as *fish* by using analogy.

A more robust way to use analogy is to use only two dictionary words at a time, where the beginning is taken from the first word and the ending from the second, with maximum overlap. The French Onomastica system JUPA (JUst PAsTe) (Yvon, 1994), inspired by the work by Byrd and Chodorow (1985), uses this technique. In this system the overlap in the matching must be at least one letter, and the overlapping letters must be consistently transcribed. These restrictions are too strong in many cases, a strategy like that shown in Figure 13 must be used.

WORD	MATCH	PICK	PASTE	PICK2a	PASTE2a	PICK2b	PASTE2b
action	/a k t i f/ a c t i f s o l u t i o n /s o l u s j ɔ̃/	/a k t/ a c t t i o n /s j ɔ̃/	*a k t-s j ɔ̃/ inconsistent!	/a k t/ a c t i o n /j ɔ̃/	*/aktjɔ̃/ wrong!	/a k/ a c t i o n /s j ɔ̃/	/aksjɔ̃/ correct!

Figure 13. The behaviour of the transcription system JUPA (Yvon, 1994).

In this figure the transcription of the word *action* is obtained from the words *actif* and *solution*. First the maximum overlap is picked, but the transcription of the *č* in *actif* is not consistent with the *č* in *solution*. Therefore the overlap is reduced to zero, pick2, which gives the two transcriptions in paste2a and paste2b, where the second is the correct one. The problem is to choose the correct solution. The system also produces a score for the quality of the transcription. The score is the sum of two parts, the size of the overlap compared to the size of the unknown word; and the number of pairs of combined words that support the solution compared to the total number of words that could be paired consistently.

A problem with this approach is that the system will not output any transcription at all if the new word cannot be matched with two words in the dictionary. A large dictionary is necessary.

5.6. Morphological analyser

The multi-morphemic structure of surnames have been shown by Spiegel et al (1989) and Belhoula (1993). The name analysis in this paper showed that 63% of the surnames in Sweden were either compounds or son-names. The rest of the names are often obtained by adding prefixes or suffixes in a similar way as for English surnames, as shown by Coker et al (1993). They also included the concept of rhyming. The basic idea is that it is safer to concatenate transcribed morphs from a dictionary than to use letter-to-sound rules for the whole words. This approach is similar to analogy, but more controlled, since the selection and transcription of the morphs is controlled by humans, as is the way they are put together.

The two-level morphology Twol, designed by Koskenniemi (1983), is a method to do morphological decomposition. It has primarily been used by computational linguists to tag text with parts of speech. The method can also be used for transcription, by including transcriptions of the morphs among the parts of speech tags. The Swedish morphological analyser Swetwol, constructed by Karlsson (1990), consists of a lexicon with 45,000 core vocabulary items. The bulk of the words were derived from Svenska Akademiens Ordlista (Svenska Akademien, 1990) and a set of eight two-level rules. SWETWOL is based on classical Swedish grammar and can form words by inflection, derivation and compounding. It consists of about 300 mini-lexicons, in which each item must point either to another mini-lexicon or to the end_of_word symbol *ě#í*. The system uses the 12 basic parts of speech: **N, A, PRON, ABBR, NUM, V, AD-A, ADV, INTJ, KONJ, PREP** and **INFMARK**.

All the morphs in the dictionary have been transcribed with a semi-automatic procedure, using the KTH text-to-speech system (Magnusson et al., 1990). The transcriptions are included among the parts of speech tags between slashes.

Analysis of inflection, derivation and compounding is done in the following manner: There may exist a number of top lexicons for different applications, like non-name words, place names and surnames, as well as mini-lexicons for inflectional and derivational categories. A lexical search consists in the establishing of a path through a series of lexicons. A lexical search is initiated in a top lexicon, e.g., MAJORS. Each lexical entry contains three fields:

orthographic field	link field	tag field
the orthographic strings and other symbols	links to mini-lexicons	the parts of speech tags and transcriptions

In each lexicon searched, a search is successful if the whole (of the remaining) string, or a substring of it starting from the left, is found as an orthographic entry. If a search in one lexicon is successful, the contents of the link field points to the lexicon that will be searched next. This next lexicon will then be searched for a match of the remainder of the input string. The search comes to an end, either if the remainder of the input string is empty *and* the next lexicon has an end-of-string mark, *ě#í*, as an entry (a successful search), or if no match could be found (a failed search).

If the search was successful, the contents of each of the lexicon entries that were hit will be concatenated in two separate streams, an orthographic stream and a tag stream. In the orthographic stream a morpheme boundary mark is inserted before each new addition. There may be several successful searches, which will then be passed on as alternative analyses.

An example of how a derived word form is analysed is shown in Figure 14. The adjective *kunglig* (ëroyalí) is derived from the noun *kung* (ëkingí), which has a lexical entry in the MAJORS lexicon. The mini-lexicon NDER in this example contains two derivational suffixes, of which one is *lig*. The Swedish system contains 12 such derivational suffixes, as well as 30 derivational prefixes. Finally the end of word symbol is found in the mini-lexicon NOMGEN.

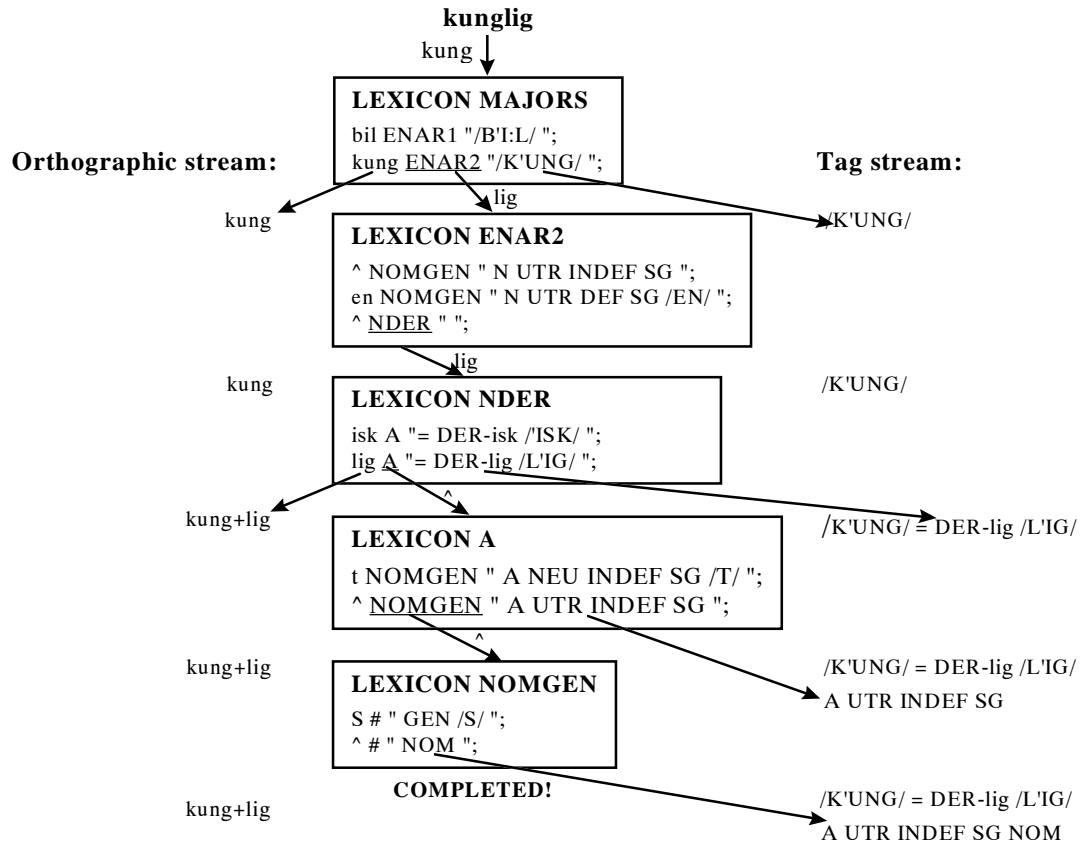


Figure 14. The morphological analysis of the derived word form *ëkungligí*.

The analysis of an inflected word form is performed as shown in Figure 15.

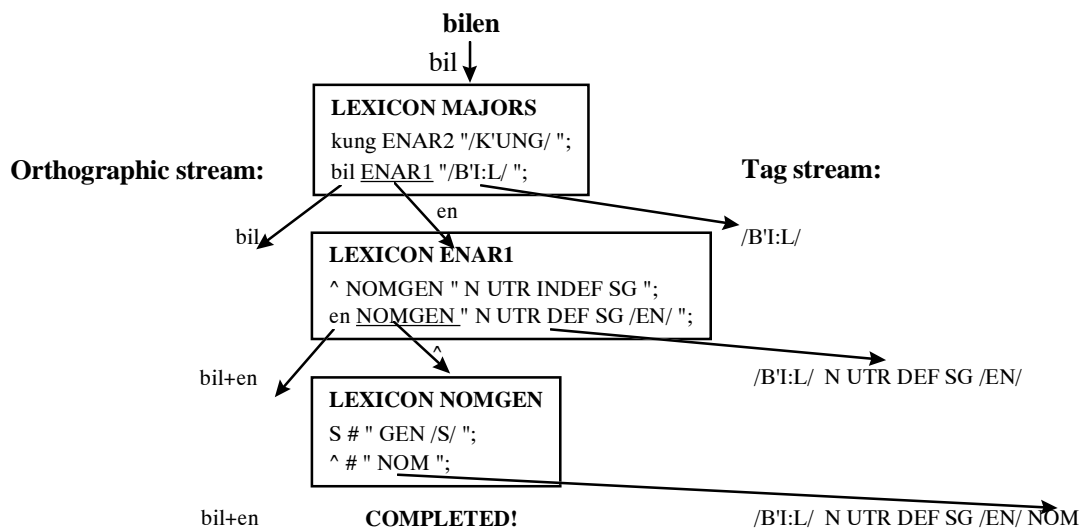


Figure 15. The morphological analysis of the inflected word form *ëbilení*.

The stem of the word *bilen* (ëthe carí), *bil*, has a lexical entry in the MAJORS lexicon. The ending *en* is found in the mini-lexicon ENAR1 and the end of word symbol is again found in the NOMGEN lexicon. Compounds are obtained by linking an entry to the lexicon **ROOT** which in turn links the entry to the **MAJORS** and **COMPOUNDCODAS** lexicons. **MAJORS** is the lexicon containing the root morphs of nouns, verbs and adjectives; **COMPOUNDCODAS** contains those words that only occur in the second part of compounds, e.g. *-aktig* in *blÅaktig* (ëbluishí).

SWETWOL has eight two-level rules that are compiled into a run-time finite-state automaton. These rules are used when the lexical and the graphemic surface representations differ. The rules have the format shown in Table 23.

Table 23. The format of the two-level rules (Trost , 1993).

Target	Operation	Context
lexical:surface	<= the context gives the target => the target lexical:surface pair exists only in this context <=> the target pair only exist in this context and vice versa /<= the target lexical:surface pair can be missing in this context	[] [] where the left and right context is separated by ë_í. The context is formulated with regular expressions, e.g., lexical:surface lexical:surface

The regular expressions used in the rules are listed in Table 24.

Table 24. The regular expressions used in the two-level rules (Trost , 1993).

Operation	Regular expression	Explanation
concatenation	a:b c:d	a:b followed by c:d
alternation	a:b c:d	either a:b or c:d
conjunction	a:b&c:d	a:b and c:d
Kleene star	a:b*	zero or more a:b
Kleene plus	a:b+	one or more a:b
negation	@V	if anything then not V
term negation	\V	an element, but not V
difference	V-e	any V except e
ignoring	V*/a:b	zero or more V, ignoring any a:b
containing	#{a:b c:d}	any string that contains at least one a:b or c:d

Some examples of rules used in the Swedish system are listed in Table 25 along with some short explanations and examples.

Table 25. Four of the eight Swedish two-level rules (Karlsson, 1990).

Swedish two-level rules	Explanations
D:t <=> :t ;	D is realised as t before t, else as d, e.g. <i>lett</i>
m:0 <=> :m_N: ;	m is deleted between m and N, e.g. <i>simning</i>
Z:s => #::*_#::0* :X ;	enables lexical Z to be realised as s when there is at least one compound boundary to the left, e.g. allows <i>skolbokshylla</i> but not <i>skolsbokshylla</i>
%: /<= %::*_ ;	constrains compound formation. The lexical segment % occurs in connection with finite verb forms. This rule prohibits the occurrence of more than one % in the same word

The system will output a number of possible solutions. The simplest way to pick the correct one is to choose the one with the least number of morphs, since small words often can be compounded and found in larger words. The parts of speech tags can be parsed in order to select the most probable solution. The transcription is then obtained by gluing the transcribed morphs together with a small set of rules.

5.7. The Onomastica results

The Onomastica partners have used different techniques for grapheme-to-phoneme conversion for names. The results found in these reports are summarised in Table 26.

Table 26. The word error rate for surnames transcribed with different approaches.

Language	Word error rate (best reported result)	Approach used for best result
Danish	34%	Table-lookup
Dutch	52%	Rules
English	33%	Rules
French	10%	Analogy
German	23%	Rules
Greek	11%	Rules
Italian	2%	Analogy
Norwegian	34%	Table-lookup
Portuguese	7%	Rules
Spanish	2.2%	Rules
Swedish	7%	Morphology

These are not the final results, but the differences in performance and techniques chosen might reflect the complexity of the task for the different languages. For some languages, like Spanish and Italian, the spelling is close to the pronunciation, while the relations between spelling and pronunciation are quite complex in languages like Danish, Norwegian, Dutch and English.

5.8. Summary

There is no universal single best method to do the grapheme-to-phoneme conversion. It depends on the language and on the available corpus. Generally the context dependent rules seem to work quite well. The statistical methods like neural networks and HMMs need a large dictionary and work best with regular words. The analogy approach is plausible for languages where the spelling is close to the pronunciation, like Spanish and Italian. Morphology is preferable in languages with a lot of compounds and a regular morphology, like German and Swedish. The best general solution seems to be to use more complex system:

1. look up the word in an exceptions dictionary
2. do language identification on the word that could not be found in the dictionary
3. do look-up of morphs dictionaries followed by morphological decomposition
4. finally apply rules to the remaining names.

If the name transcription system is to be used in a text-to-speech system, the global context can be used as well. This can for example be done by using parts of speech tags obtained by a parser to select the correct pronunciations for hyponyms; or by applying co-articulation across word boundaries.

6. Description of the transcription system

The KTH text-to-speech system is a multi-lingual system that covers about 10 languages (Carlson et al., 1991). The grapheme-to-phoneme conversion is primarily done using context dependent rules (Carlson and Granström, 1976). The system has been modified to cope with proper names in a previous project (Carlson, et al., 1989,1990). Further modifications of the context dependent rules and the extension of name to the morphology approach (Magnusson et al., 1990), have been done for the Onomastica Project (Gustafson 1994,1995). Figure 16 shows the layout of the present grapheme-to-phoneme system for surnames.

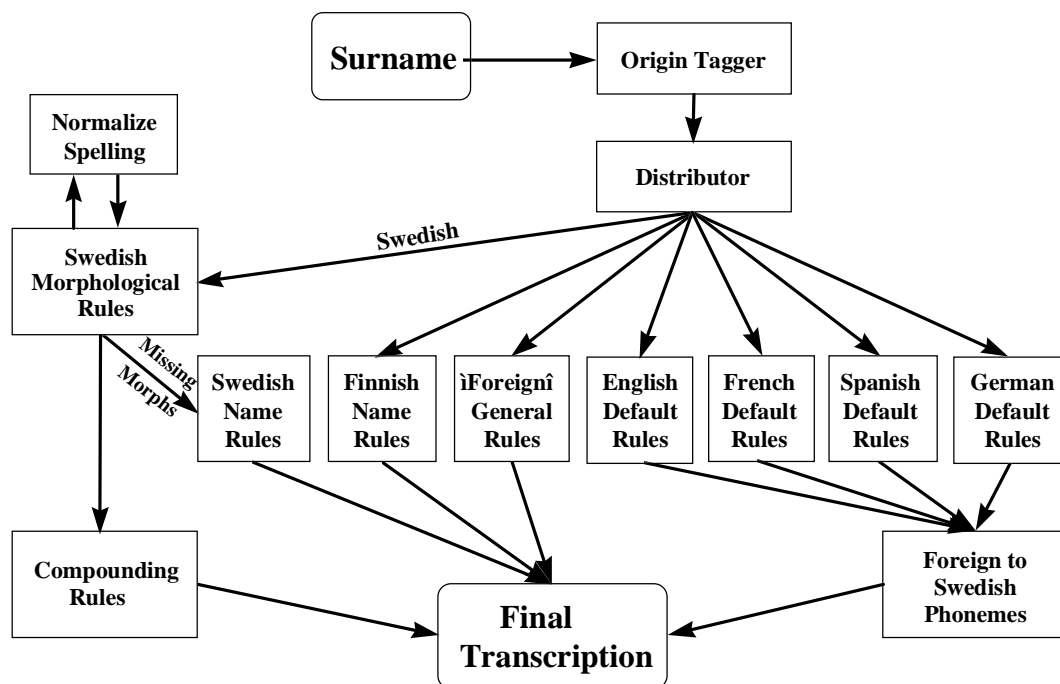


Figure 16. The KTH grapheme-to-phoneme system for surnames.

First the origin of the name is determined to simplify the automatic transcription. For more details on the origin tagger, see Chapter 7.1. 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 morphological analyser. Morphs with stress and boundary markers are obtained from the morphological analyser. These are merged into complete transcriptions using context dependent transformation rules. The names that the morphological analyser could not recognise are normalised in spelling and then processed again. The names that the analyser still cannot manage are transcribed with the ordinary Swedish grapheme-to-phoneme rules adjusted for names. The foreign names are first run through language specific grapheme-to-phoneme rules, with language specific phonemes. Finally these phonemes are mapped to the closest Swedish equivalents. The sounds are generated with a version of the KTH formant synthesiser OVE III, described by Liljencrants (1968).

The grapheme-to-phoneme part of the KTH text-to-speech system will be described in this chapter.

6.1. Grapheme-to-phoneme conversion rules

The grapheme-to-phoneme rules used in the KTH text-to-speech system for Swedish were updated with name-specific rules. The original system for Swedish had 450 rules. Of these, 250 were kept in the rule system for names. These rules were general rules of which some are exemplified in Table 27.

Table 27. Some rules in the notation used in the KTH system, RULSYS, that are used for transcription of both non-name words and names.

Example of rules	Description
SCH → SJ	Mapping the spelling SCH to the phoneme
LCRULE W → V	Mapping W to V
<KONS,ANT,COR> → <-ANT> / <KONS,-ANT,COR,-TENSE> _	Retroflexions of coronal consonants after R and retroflex consonants
R → / _ <KONS,-ANT,COR>	Removing of R preceding retroflex consonants
<D,-ANT> → 2D	Mapping D with feature <-ANT> to 2D

Most of the removed rules were of a morphological nature, that inserted morph boundaries between certain known morphs. For example:

G Y M N A S I E → J Y M N <A,STRESS,1STRESS> S I E # / _ <SEG> *gymnasieskola*
 → # / <WORD> E,(1) M O T & <SEG> <SEG>(2,2) *emotse*
 → # / <WORD> U N D E R & <SEG> *underhÅlla*
 → # a / & A K T I G <-SEG,FB> *l'gnaktig*

About 500 name rules have been inserted in the rule system for names. Many of these rules insert a morph boundary between known name-morphs, while others adjust the stress pattern according to the ending. For example:

S K I → SJ / _ ÷ L D *Bergski`ld*
 → # / _ F E L D,(1) T sw *Bergfeldt*
 I → <STRESS,1STRESS> / _ N <NAME,WORD> *Wallin*
 <VOK> → <-STRESS,-TENSE> / _ <KONS>(,) I N <NAME,WORD> *Wallin*

The stress adjustment rules give two-syllable son-names accent I and those with more syllables accent II. They also move the primary stress to a later syllable in names ending with *-in*, *-elius* and other such endings. The name rules were hand-coded using the transcriptions from the previous project. This approach is quite time consuming and the result is not satisfactory, as can be seen in the evaluation in chapter 8.

Rules for several languages are used in the name pronunciation system. Names that have been tagged as English will be transcribed by the English grapheme-to-phoneme rules. The transcriptions obtained are then mapped to the closest allophone used in the Swedish Text-to-Speech system. Some of these mappings are shown in Table 28.

Table 28. Some mapping between English and Swedish phonemes.

English	w	ð	dʒ	ʃ	θ
Swedish	v	d	dʃ	ʂ	t

6.2. Morphological analysis

The morphological analyser described in chapter 5.6 has been updated to cope with names as well. The current morphological lexicon, listed in Table 29, 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 book, compiled during a previous project (Carlson et al., 1989).

Most of the work has been done on the surnames since they are often made of more than one morph, taken from a set of surname morphs. The street names are even easier to cope with, since the use ordinary words compounded with *-gatan* (‘street’) or *-vägen* (‘road’). The only change done for street names was to reduce the restrictions for compounding.

Table 29. The updated Swedish morphological lexicon.

Group of morphs	Example	number of morphs
ordinary Swedish morphs	<i>hopp</i>	45,000
place names	<i>stockholm</i>	4,361
full surnames from old lexicon	<i>olson</i>	24,083
first names, uncategorised	<i>prahl</i>	1,570
female first names	<i>marie</i>	2,243
male first names	<i>anders</i>	1,931
initial only, compound forming surname morphs	<i>wahl</i>	1,877
compound forming surname morphs	<i>berg</i>	554
stress-taking surname morphs	<i>elius</i>	75
non-stress taking surname morphs	<i>ner</i>	120

There are restrictions in the use of the name morphs. This was demonstrated by a governmental surname committee in 1964 (Statens offentliga utredningar, 1964). They constructed new family names by combining 2,200 initial morphs with 230 end morphs. They decided that disyllabic end-morph only could be combined with monosyllabic initial morphs, in order to avoid generating inappropriate names like *Lindenhagen*. About 100 of the initial morphs should only be combined with stressed end-morphs, e.g., *Ygn-ell*, while 300 could only be combined with unstressed end-morphs, e.g., *Ab-man*. The rest of the 1,800 initial morphs could be combined with any of the remaining 130 end-morphs, with some exceptions: no combinations of morphs are allowed that would result in a coming together of the same vowels or consonant clusters, e.g., *Ask-skog* and *Grane-ong*. In the telephone directory only a few names of this type are found, e.g., *West-stedt* and *Lilje-ong*. There are also restrictions to the effect that the morphs should not include identical initial or final consonant clusters, e.g., *Blom-blad*, *Lund-lind*. Again there are a few examples in the telephone directory violating this rule, e.g., *Kvile-kval*, *Stocken-stam* and indeed *Lind-lund*.

Table 30 lists the ten most common initial morphs and the ten most common end morphs. The table shows the number of subscribers with each possible combination. In addition to the illegal combination of identical morphs (indicated by a star), there are some combinations that are rare or even non-existent, such as *Holm-mark*, *Sand-dahl*.

Table 30. Number of subscribers with names that are combinations of the ten most common initial morphs and the ten most common end morphs

	berg	str [^] m	gren	lund	kvist	holm	dahl	stedt	mark	wall
Lind	13,781	13,056	12,024	1	12,009	4,061	6	803	1541	2,041
Lund	11,233	6,345	10,822	*	100,504	1,097	995	692	2,722	759
Berg	*	10,882	5,662	10,202	6,884	265	965	624	418	1,216
Sj[^]	8,319	3,771	5,255	1,442	1,853	1,742	459	1,335	47	311
Ek	2,360	5,176	21	7,631	57	1,594	1,250	345	148	976
Ny	6,525	7,083	3,744	928	2,011	557	243	395	84	69
Holm	7,076	3,773	5,260	945	4,601	*	191	461	0	73
S[^]der	6,453	4,407	1,206	3,725	2,027	1,382	127	168	154	130
Sand	8,937	5,444	1,501	284	1,281	141	2	330	97	193
Eng	1,562	7,955	37	3,956	894	239	1434	104	26	1,126

Many of these morphs have been constructed by adding *-e*, *-en*, *-er*, *-a* 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. Table 31 shows the number of morphs with these suffixes for surnames and place names. If these suffixes are removed from the surname morphs the number of different morphs in the first part decreases from 3,990 to 2,490 and in the second part from 595 to 476. If these suffixes are removed from the place name morphs, the number of different morphs in the first part decreases from 2,546 to 1,989 and in the second part from 556 to 433.

Table 31. Number of two-morphemic names with morphs ending with certain suffixes.

Ending	Surnames		Place names	
	First	Second	First	Second
-E	933	60	468	88
-EN	358	31	28	39
-ER	374	45	71	14
-A	26	16	244	168

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 75 morphs in the lexicon that are always in final position and get the primary stress regardless of the first part, e.g., *-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 whether the names actually exist in the Swedish telephone book. Only morphs that in the telephone directory occur with the *-er*, *-en*, *-e* endings are allowed to be combined with these, even though some others could theoretically 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 Figures 17, 18 and 19, decreases with rank. The last 90,000 surnames, for example, only occurred once in the database and most of these were either foreign or misspelled.

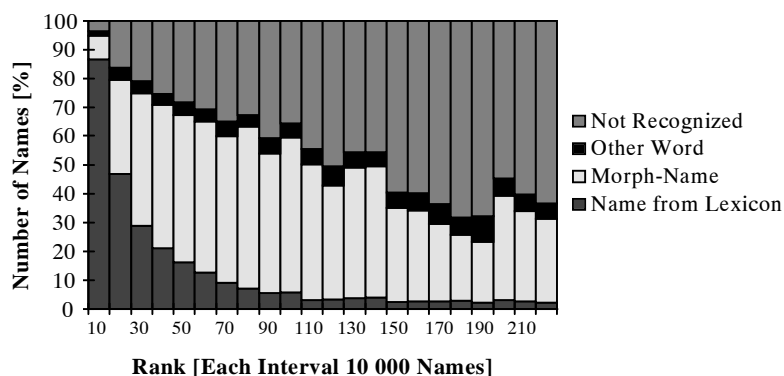


Figure 17. The coverage of the Twol analyser for surnames.

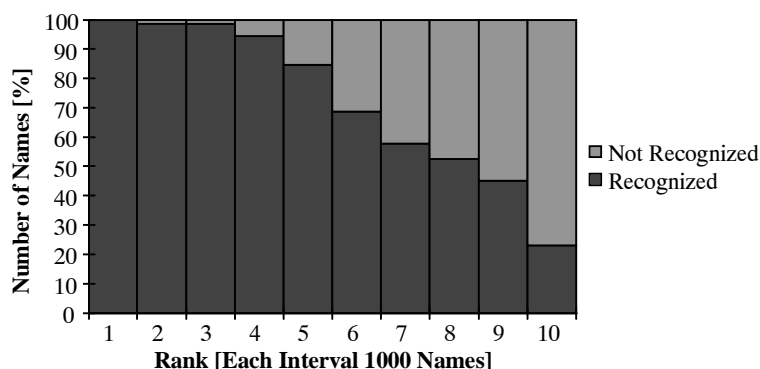


Figure 18. The coverage of the Twol analyser for first names.

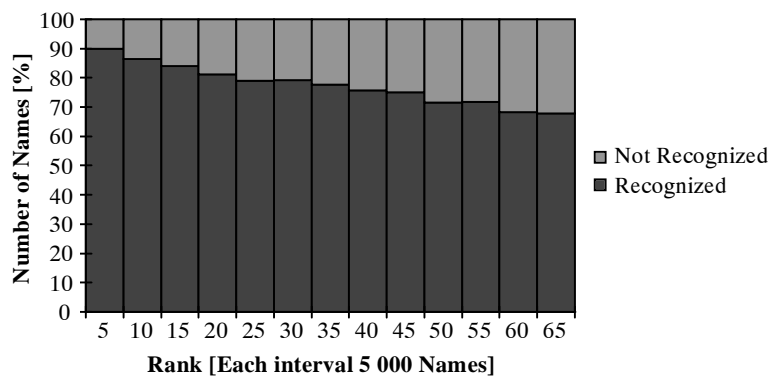


Figure 19. The coverage of the Twol analyser for street names.

The analogy approach that was described in chapter 5.5 is quite sufficient when dealing with new names sharing the same patterns as other known names. One way of increasing the coverage of the system is therefore to include an analogy component. This has been implemented for surnames. All transcribed surnames with verified transcriptions and tagged as being Swedish have been analysed. These 38,000 names have 102 different consonant clusters with 123 different transcriptions. See Figures a5 and a6 in the Appendix for a graph illustrating these clusters. The legal initial consonant clusters in Swedish words have been described by Sigurd (1967). His model can generate 46 phoneme clusters of which 4 are disregarded as not legal. One of these, *ēsri*, does occur in surnames. Another group of consonant clusters that are not

found in Swedish words comes from surnames of German origin, for example *ështëri*, *ështëli*, *ështëwi*, and *ështëni*. Only two of the legal consonant clusters for Swedish words are not found in surnames, namely *ështëni* and *ështëvi*. The reason for this might be that words with these sequences are perceived as phonaesthetically negative:

fnalla (ëeat snacksí); *fnas* (ëhuskí); *fnask* (ëprostituteí); *fnissa* (ëgiggleí);
fnoskig (ëdottyí); *fnurra* (ëtangleí); *fnyk* (ërubbishí); *fnysa* (ësnortí);
skvala (ëgushií); *skvaller* (ëgossipí); *skvalp* (ësplashí); *skvatter* (ëgabbleí); *skvimpa*
 (ëspillí); *skv%ottd*(ësquirtí)

The trigraphs that were found in surnames have been included as transcribed prefixes in the morph lexicon. This has increased the coverage by about 5%. The transcriptions that were generated in this way have the same error rate (6%) as the other transcriptions obtained from Twol.

6.3. Normalising the spelling of names

Many names do not have a unique spelling, which contributes to the difficulties involved in transcribing them. Some of the spellings are invented by people with common names, who want to make their names more distinctive by spelling them in an unorthodox way. The use of different spellings seems to be more popular in Swedish 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 a unique reference. Swedish first names have up to 66 different spellings, see Table 32, while Italian names only have up to 6 spellings.

Table 32. Some examples of different spellings of names.

The name Persson has 21 orthographies	$\begin{array}{c} \text{SS} \\ \text{P (H) E (H) R (H) SZ O (H) N (N) (E)} \\ \text{f S Z} \end{array}$
The name Ann-Katrin has 66 orthographies	$\begin{array}{c} \text{K} \\ \text{A N N (E) (-) C A (H) T (H) (A) R I N (N) (E) (...)} \\ \text{CH} \end{array}$
The name Therese has 26 orthographies	$\begin{array}{c} \text{EE E...SS (EE)} \\ \text{T (H) E (H) R E S (E)} \\ \text{... ... Z (...)} \\ \text{» » C} \end{array}$

The different spellings do not always follow ordinary orthographic conventions. One practice that has been observed is the insertion of *ëhi*, *Bhlom*, another the use of *ëvi* instead of *ëksi* in names ending with *-son*, the name *Eriksson*, for example, has an alternative spelling *Erixson*. Some other popular replacements are: *s*→*z*, *k*→*q*, *k*→*c*, *k*→*ck*, *Å*→*aa*, *ö*→*ae*, *^*→*oe*, *^*→*eu*, *i*→*ie*, *o*→*ou*, *f*→*ph*, *v*→*fv*, *v*→*w*, *j*→*i*. The spelling of the names must be normalised in order to simplify the automatic transcription.

To cope with these strangely spelled names, a set of context dependent rules that normalised the spelling was developed. These rules, for example, convert strange spellings like *Qvarn* to more common ones like *Kvarn*. The rules were used only on the names that had been tagged as Swedish, otherwise they might corrupt the spelling of foreign names.

7. Transcribing names with foreign origin

The pronunciations of words in Swedish are often affected by the origin. The same grapheme string can be realised differently depending on the origin of the word. The initial consonant cluster *ëchí* is for example realised [ʃ] in words of French origin, like *ëchampagneí* and *ëcharmí*, and [ç] in words of English origin, like *ëchipsí* and *ëcharterí*. Names are also pronounced differently depending on the origin of the name. The name *Jones*, for example, is pronounced [j˘u:nəs] if regarded as Swedish and [dʒoʊns] if it is recognised as English. Many names have been incorporated without going through the normal linguistic processes of changing the spelling and pronunciation of borrowed words according to the native phonotactics.

Another difficulty is that names retain old spellings that have not survived the spelling reforms for ordinary words. This means that names have kept both foreign and old Swedish patterns, which makes them hard to cope with for an automatic pronunciation system.

7.1. How to deal with foreign names

The realisation of foreign words depends on a number of factors, such as the geographic and phonemic closeness of the foreign language and the type of word: only names can get an authentic pronunciation (Elert, 1971). Furthermore it depends on the time it has been used in the new language, e.i., to what amount it has been incorporated in the native language system. The pronunciation of names also depends on the strategy used. It is the policy of the news on Swedish Television to always pronounce names the way the bearer wants it to be pronounced. When a new president is elected somewhere in the world, they try to get in contact with someone who has contact with the president, in order to get the correct pronunciation of the name. The opposite strategy is often used in France where all names occurring in France are regarded as French, and consequently pronounced with French rules. Some factors that influence the realisation of a foreign name have been listed by Mengel (1993) in the Onomastica project:

- 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 when pronouncing a foreign name. These agree with the principles Elert (1971) postulated for the pronunciation of foreign words in Swedish. These are some of the principles, exemplified with transcriptions from the Onomastica database:

- 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, e.g., the Swedish name Greger [gre:gər] gets the Dutch transcription [xreɣər].
- 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 ones, e.g., the English name Winston [ˈwɪnstən] is transcribed [ˈvɪnstən] in Swedish.
 b) Enlarge the native phoneme inventory, e.g., the English phonemes [ð] and [θ] are added to the Swedish inventory to get

Heather[ˈhɛðər] and Keith[ˈkiːθ].

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,
 e.g., the Swedish town Göteborg [jø:tebøɾj], is transcribed [jɛtebøɾg] in Greek, where the Swedish way to pronounce *ēí* is known, but it has to be mapped to the closest Greek phoneme.
- b) If the realisation of the foreign grapheme is unknown map it to the closest native grapheme and transcribe according to the new spelling,
 e.g., Göteborg is mapped to Goteborg in Spanish and is transcribed [goteβoɾ].

7.2. The origin tagger

To cope with the influence of the origin of the names a set of origin tags have been introduced in the name pronunciation system. The system is designed to imitate a Swedish person attempting to pronounce a foreign name. Therefore it is not certain that the origin tags will be 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, 14 tags for origin have been included. The use of an origin tagger for a German text-to-speech system has been described by Henrich (1989)

The tagging has been done using the KTH text-to-speech system with *iphonological* rules, (Carlson et al., 1989), which recognise orthographic patterns that are specific to different languages, see Figure 20. The rules change the tag for unknown (uk) to some of the other origin tags. All names ending with *-izzi* are for example regarded as Italian and names that include the grapheme string *eschrai* are tagged as German. Since most 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.

```

uk→ ar / DEH _ <WORD>
uk→ as / CH<VOK>NG _ <WORD>
uk→ de / SCHRA<SEG>* _ <WORD>
uk→ en / GHT _ <WORD>
uk→ es / ERA _ <WORD>
uk→ fi / NEN _ <WORD>
uk→ fr / EAU<KONS>* _ <WORD>
uk→ gr / OTIS _ <WORD>
uk→ it / ZZI _ <WORD>
uk→ se / IUS _ <WORD>
uk→ sl / IC _ <WORD>

```

*Figure 20. Example of origin rules for names, where *ëukí* stands for unknown, *ëarí*-Arabic, *ëasí*-Asian, *ëdeí*-German, *ëení*-English, *ëesí*-Spanish, *ëfíí*-Finnish, *ëfrí*-French, *ëgrí*-Greek, *ëití*-Italian, *ëseí*-Swedish and *ëslí*-Slavic. Features are enclosed by <>, <WORD> stands for next word, i.e. all origin tag are added on the end of the names. <SEG> represent any phoneme. *ë*í* denotes zero or more instances.*

In the current system a new origin tagger has been introduced. It uses trigraph statistics for the 11 languages included in the Onomastica database. The language

specific databases include names of different origins. This introduces some noise in the statistics, but since most of the names are of the same origin, it does not diminish the performance of the system. The Swedish surname database contains many Finnish, Arabic and Slavic names. These were removed from the Swedish database and used for making trigraph statistics for these languages. Then the statistics were used to detect names of these origins in the other databases as well. These names were removed from all databases, giving lexicons of 17,058 Arabic, 16,335 Finnish and 41,507 Slavic names. In order to reduce the noise, the statistics for the other languages were re-computed without these names. For each trigraph the number of occurrences of that trigraph is divided by the total number of trigraphs in each language. This is the probability of the trigraphs occurring in a name of that origin. The trigraphs and their probabilities are stored in language specific dictionaries. The 50 most common trigraphs of each language are presented in Tables a34-a37 in the Appendix. The origin of a name is obtained by the following algorithm: for each language, compute the sum of the probabilities for the trigraphs included in the name, then pick the language with the lowest score. Trigraphs that can not be found in the dictionary will get a penalty score.

If this approach were to be used on the Swedish telephone book the score for the Swedish name would be lowered, since most names in Sweden are recognised as Swedish. An example are the many surnames in Sweden that have a German origin. These names get a slightly higher German score than Swedish. To be able to recognise them as Swedish, as most people do, the Swedish scores have been lowered.

The accuracy of the two methods has been computed for 228,000 surnames in the Swedish database. The origin tags for this database have been obtained by using the two different systems to give the initial tags. If the systems gave different tags, one of them was selected by various rules. The database was then manually corrected by scanning the dictionary in two phases. In the first phase the names were ordered alphabetically from the beginning, and in the second phase they were ordered from the end of the words. This was quite useful since the endings are specific for different languages. The result of this manually correct origin tagging is presented in Table 33. It indicates the different influences on Swedish names as well as the number of immigrants from different areas.

Table 33. The probable origin of the surnames of the 4.1 million subscribers in the Swedish telephone directory. Swedish, Finnish, German, Slavic, Arabic, Spanish, French, Dutch, Italian, English, Portuguese, Asian and Greek.

Probable origin	Se	Fi	Ge	Sl	Ar	Sp	Fr	Du	It	En	Pt	As	Gr
Number of occurrences (k)	382	89	46	44	29	24	18	18	18	14	11	4	3

The database contained 63% Swedish names that covered 92% of the subscribers. In Sweden approximately 20% of the inhabitants are foreigners. Many of them are Finnish, which is reflected in the table. A reason why only about 8% of the subscribers have names that are tagged as foreign might be that most of the immigrants do not have a telephone.

The rule-based approach predicted the correct origin in 68% of the names, covering 94% of the subscribers. This is an improvement of only 5% from the initial guess that all were Swedish. The probabilistic method using trigraphs worked better. It predicted the correct origin in 95% of the names, covering 98.5% of the subscribers. Since the

correction was done by scanning the 228,000 names, these figures are not exact, but the real result is probably of that order.

7.3. Comparison of first names in five languages

To exemplify the problems of transcribing names with foreign origins, the databases containing first names from Great Britain (16,111 transcribed first names), France (12,383), Germany (31,979), Italy (35,013) and Sweden (10,461) were examined. The difference in size of the databases could be adjusted by selecting the 10,000 most frequent names, but since the frequency was only available in the Swedish database the databases could not be truncated. Therefore, 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 34 shows the average number of letters and phonemes in the first names and in the 10,000 most frequent common words.

Table 34. Average number of letters and phonemes in First Names (FN) and Non-Name Words (NNW) in five languages.

	Letters in FN	Letters in NNW	Phonemes in FN	Phonemes in NNW
Sw	7.3	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	9.0	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. Since broad transcriptions were used the actual realisation of individual phonetic symbols varied from language to language.

The most common phonemes in each language's first names are shown in Table 35. The table shows that the most common phoneme is [a] in all languages, except for English where it is the [ə].

Table 35. The most common phonemes in the transcriptions of first names, with occurrence in percent under each phoneme.

Sw	a 12	l 7	r 7	n 7	ɪ 7	s 6	k 6	e 4	t 4	j 4
En	ə 9	ɪ 8	n 8	æ 7	r 6	l 5	i 4	s 4	ε 4	m 4
Fr	a 13	i 10	ʀ 8	l 7	n 6	ε 5	m 5	s 4	d 4	t 4
Ge	a 10	t 7	n 7	r 6	l 6	ɪ 5	ə 5	k 4	ʋ 4	s 4
It	a 15	o 11	e 9	i 8	n 8	r 8	l 6	t 5	m 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 have the largest number of phonemes (about 45), but the smallest number of phonemes

per name (about 5.5). If from each country you pick the name that contains as many as possible of these phonemes you get the following names:

Sw	Nils-Einar	[nɪlsejnar]
En	Alexander	[æliɣzændər]
Fr	Alexandrine-marthe	[aleksāndrinmart]
Ge	Weichselgortner	[vaiksəlgertnə]
It	Vittorio-emanuele	[vitorjoemanuele]

The databases contain altogether 88,000 different names of which 79,000 only occurred in one country while 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 36 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 (the letter *ġ* does not exist in ordinary Italian words).

Table 36. 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	-

Some preliminary statistics on the lengths of first names and surnames in ten of the languages in the Onomastica project are shown in Figures a19 and a20 in the Appendix.

7.4. Pronunciation of an initial *ġ* in different languages

When examining the Onomastica databases it was noticed that the letter *ġ* in initial position got quite different transcriptions in the different languages. Some examples of transcriptions of the same names in different languages are shown in Table 37.

Table 37. The pronunciation of an initial *ġ* in some first names of different origins.

Language	Joyce	Jacqueline	Juan	Joakim
German	'dʒɔys	ʒak.li:n	'χu:an	'jo:a.kim
Danish	dʒʌjs	ʃaɟl'inən	hu'an	ʃ'o:akim
Spanish	not found	xa.'kel	'xwan	xo.a.'kin
French	ʒɔjs	ʒa.kə.lin	xwan	jo.a.kim
Italian	'dʒɔis	dʒak.'lin	'hwan	'jo.a.kim
Dutch	dʒɔjs	ʒa.kə.'li.nə	xu.'wan	'jo.wa.kim
Norwegian	'jɔys	ʃak'li:n	'hu.ɑn	'ju:.ɑ.kim
Portuguese	'ʒɔj.sə	ʒɐ.kə.'li.nə	ʒw.'ẽ	ʒw.ɐ.'ki
Swedish	'dʒɔjs	ʂak'lin	ʃu'an	˘ju:akim
English	dʒɔis	'zakəli:n	hwan	joʊ'akim

The English names are mostly transcribed with [j] in Swedish, but in some cases the [dj] cluster has been used to imitate the English [dʒ]. Spanish names like *Juan* [xwan] has been transcribed with [x] in French and Dutch, [h] in English, Danish, Norwegian and Italian, [ʒ] in Portuguese, [χ] in German and finally with [ʃ] in Swedish.

The different pronunciations seem to be dependent on the likely origin of the name. The names that are considered to be of a certain origin get the pronunciation of $\text{ɛ}j\acute{\text{i}}$ that is most common in that language; or is mapped to the closest one in the native language, see Table 38.

Table 38. The pronunciation of an initial $\text{ɛ}j\acute{\text{i}}$ in first names of the Onomastica languages. The second row for each language shows the number of occurrences, the third the likely origins of the names, where $\text{eas}\acute{\text{i}}$ is Asian. The $\text{ɛ}-\text{ɛ}$ indicates that the language of that column is in the native column, $\text{ɛ}-\text{ɛ}$ indicates that there are no native names with initial $\text{ɛ}j\acute{\text{i}}$ in Italian.

Language		Native	English	French	Spanish	German	Asian
German	phoneme	j	dʒ	ʒ	χ	<-	tʃ
	names	1148	29	26	4	<-	126
	origin	ge	en	fr	sp	<-	as
Danish	phoneme	j	dʒ	ʃ	h		
	names	1036	85	62	4		
	origin	da	en	fr	sp		
Spanish	phoneme	x			<-		
	names	112			<-		
	origin	sp			<-		
French	phoneme	ʒ	dʒ	<-	x	j	tʃ
	names	932	52	<-	15	21	1
	origin	fr	en	<-	sp	ge, sw	as
Italian	phoneme	-	dʒ		h		j
	names	-	173		47		337
	origin	-	en,fr		sp		foreign
Dutch	phoneme	j	dʒ	ʒ	x		ʃ
	names	1026	22	38	10		2
	origin	nl	en	fr	sp		as
Norwegian	phoneme	j	dʒ	ʃ			
	names	372	8	20			
	origin	no	en	fr			
Portuguese	phoneme	ʒ	dʒ			j	
	names	465	1			1	
	origin	pt	en			ge	
Swedish	phoneme	j	dʒ	ʃ	ʃ		
	names	470	4	11	17		
	origin	sw	en	fr	sp		
English	phoneme	dʒ	<-	ʒ	h	j	
	names	197	<-	9	5	2	
	origin	en	<-	fr	sp	ge, sw	

The same mappings for $\text{ɛ}j\acute{\text{i}}$ are also found in surnames. Some of the surnames that occurred 7-10 times in the ten languages were also examined. Table a44 in the Appendix shows that many of the languages use different pronunciations of an initial $\text{ɛ}j\acute{\text{i}}$ depending on the origin. The only language that never mapped $\text{ɛ}j\acute{\text{i}}$ according to origin was Spanish where the pronunciation is always [x]. In Italian all names are mapped, since they do not have the letter $\text{ɛ}j\acute{\text{i}}$. The table also shows the origin tags for these names. These tags were obtained from the origin tagger described earlier. Names

that were recognised as English both by the origin tagger and the author were often mapped, while those recognised as Swedish and Danish were not. This supports the hypothesis that the names of known origins are influenced by the origin. Names from less known languages, e.g., Swedish, or names of unknown origin, e.g., *Just*, or of Biblical names, e.g., *Job*, are mostly pronounced as native names.

8. Evaluation of the name pronunciation system

All the automatically transcribed names, of band I, in the Swedish part of the Onomastica project were manually corrected. Only one person did this in order to obtain consistency. The correction was done in two steps: first the transcriptions were checked automatically with software that detected inconsistencies or illegal transcriptions. All transcriptions in band I were then *iproof listened*, using 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 (Gustafson, 1994).

The automatically generated transcriptions were compared to the manually corrected versions in order to evaluate the system performance. The surnames were selected as test corpus in this evaluation for two main reasons: first the task introduces various interesting problems, such as dependency on origin and ambiguous pronunciations. Second, evaluations of transcription systems for surnames have been reported in other papers, enabling a comparison of different approaches and how the language may influence the complexity of the task. The first names and place names were not used since most of them had been manually transcribed. The task of transcribing street names was not regarded to be complex enough to give an interesting evaluation.

8.1. Evaluation of three test samples

Three test samples of 1,000 surnames each were selected as shown in Figure 21. These names were transcribed by the system and the transcriptions were compared to the manually corrected lexicon. In cases where the transcriptions differed an acceptance check was done. If the transcriptions were regarded as being equally acceptable it was not counted as an error in the evaluation.

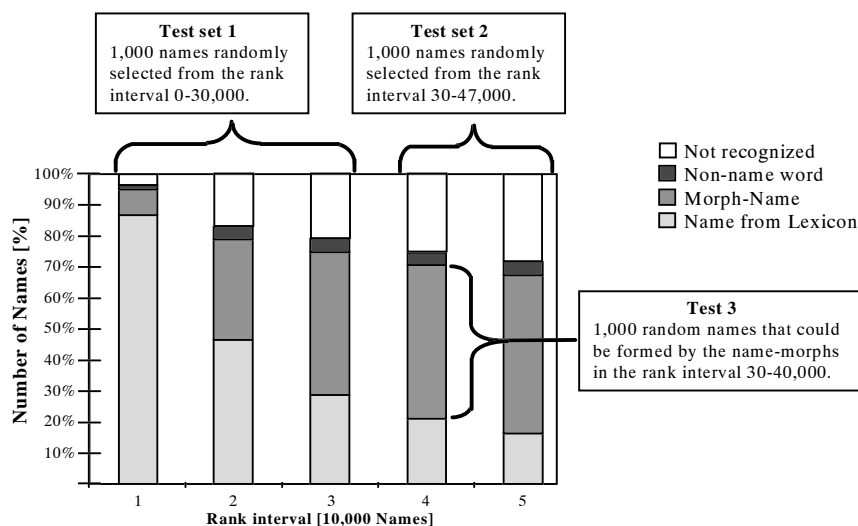


Figure 21. The three examined test samples of 1,000 surnames. The bars in the figure represent 10,000 names each, and they show the distribution of morphological analyses obtained in each rank interval.

The first test set of 1,000 names was randomly selected from the first 30,000 surnames. The system generated transcriptions with a total error rate of 3% for these

names. About 90% of the words were processed by the morphological analyser that produced transcriptions with an error rate of only 1.7%. This is not surprising since 80% of the transcriptions generated by Twol were obtained from the internal lexicon. If the lexical lookup transcriptions are separated from the generated ones the error rates shown in Table 39 are obtained.

Table 39. The efficiency of the transcription system in the first, low rank, test sample.

Transcription Source	Number of Names	Error Rate
Lexicon in Twol	670	0.15%
Twol generated	236	6%
Rule generated	94	15%
Total	1,000	3%

Only one of the transcriptions from the lexicon in Twol was considered not correct. However, this was a foreign name, *Grahovac*, which had the transcription [gr'a:hɔvatç] in the corrected lexicon and [gr'a:huvats] in the Twol lexicon. Both pronunciations are probably acceptable by a Swedish speaker. The error rate for the transcriptions generated by the letter-to-sound rules was 15%. Most of the names that could not be processed by the morphological analyser were of foreign origin.

The second test set of 1,000 names was then selected from the rank interval 30-47,000. In this interval only about 60% of the names were processed by the morphological analyser. The total error rate increased to 15%, but this can be explained by the larger number of foreign names. As can be seen in Table 40 the error rate for the Twol approach in this interval has also increased. This is probably due to the fact that only a fourth of the transcriptions generated by Twol in this test sample were taken from the lexicon.

Table 40. The efficiency of the system in the second, high rank, test sample.

Transcription Source	Number of Names	Error Rate
Twol	580	5%
Swedish Rules	177	24%
Foreign Rules	243	32%
Total	1,000	15%

The names that were not processed by Twol were passed on to the origin tagger and then sent to the appropriate rule system. As can be seen in the table, most of the names that were not processed by Twol were considered as being of foreign origin. The Swedish rules had a higher error rate for the names in this test sample with high rank names than in the first sample. This is because the name specific rules have been designed by examining the 20,000 most common names in the Stockholm telephone directory of 1988. The names in the present test set did not have the same patterns as those used to develop the rules.

The first two test samples included both Swedish and foreign names that occurred in the lexicon and were transcribed by either Twol or the letter-to-sound rules. To be able to study the quality of the transcriptions of Swedish names that had been produced with the morphology approach a third test set of 1,000 names was selected. Names were randomly selected from the rank interval 30-40,000, among those that could be formed by the name-morphs, i.e., had not been taken from the Twol lexicon

of complete names. These names were then processed by Twol, letter-to-sound rules for names and letter-to-sound rules for common words. The error rates of the transcriptions produced by the different approaches are shown in Table 41.

Table 41. The efficiency of three different approaches on the third test sample.

Approach	Error Rate
Common Words Rules	66%
Name Rules	52%
Twol	6%

The transcriptions from the rules that were adjusted for names were better than the ones for common words, but still much worse than the ones from Twol. The reason for the high error rate for the rules is that the names in the third test have different morphemic patterns than the names used to develop the rules. The errors produced by Twol were mostly of a morphological nature, such as a missing morph boundary, or a missing stress mark in the end-morph, resulting in wrong word accent. Most of these errors can be avoided by tuning the morph lexicon and the rules that merge the transcribed morphs together into complete transcriptions.

A listening test was performed to analyse how serious the transcription errors would be if they were used in a text-to-speech system. The Twol system had produced 60 transcriptions that were regarded as not correct by the author. For 35 of these, the name rules had made a different, but still not correct transcription. These names, with a total of 3x35 different transcription (Twol generated, rule generated and hand corrected transcriptions), were synthesised and presented to 14 test subjects. The subjects were all speech researchers that were familiar with the speech synthesis. This minimised the possibility of their evaluating the quality of the speech synthesis instead of the quality of the transcriptions. The test stimuli were presented to the subjects on a WWW-page with a listing of the 35 names. Each name was followed by links to the three speech files, labelled test1, test2 and test3. The order of the speech files was randomised in three different ways in order to make the test independent of the order of the different transcriptions. The subjects could listen to the test files by clicking on the links. They could listen to them in any order and as many times as they wished, before deciding on a score for the transcription. The scores ranged from 1 to 5, corresponding to:

- 1 wrong pronunciation, almost unintelligible
- 2 unacceptable pronunciation
- 3 acceptable pronunciation
- 4 correct pronunciation, but not the way I would say it
- 5 the way I would pronounce the name

The general result of the listening test is shown in Figure 22, where the scores 1 and 2 are pooled and labelled as being *ënot okí*, 3 is *ëokí* and 4 and 5 are labelled as being *ëcorrectí*. In the figure each bar contains all (14x35) judgements given by the 14 subjects for the transcriptions from each of the three approaches .

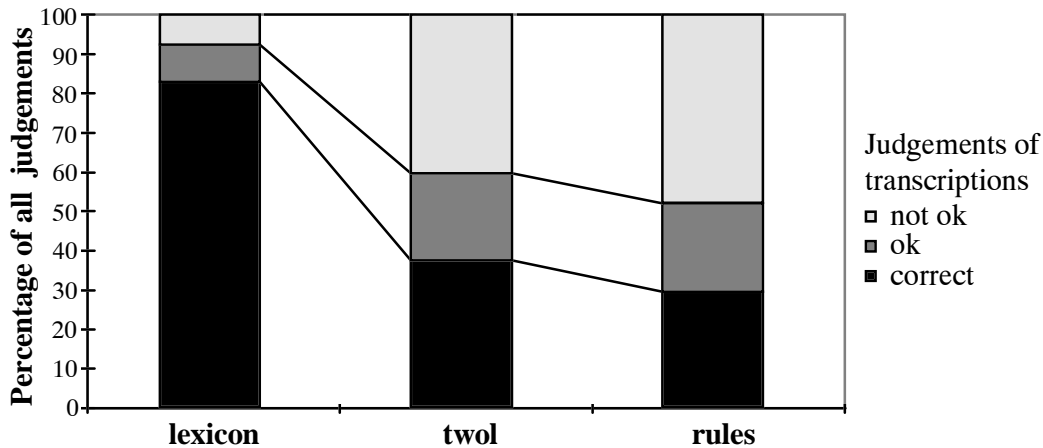


Figure 22. The overall result of the 14 subjects judgements of 35 names transcribed by Twol, rules or with transcriptions taken from the lexicon.

In the selected sample the corrected lexicon has a total acceptance of 93%, compared to 62% for the erroneous Twol transcriptions and 55% for the rules. The names that were selected had occurred only 6-8 times in the Swedish telephone directory, and they were consequently considered unusual by the subjects. Some of them were probably of foreign origin, making it even harder to transcribe them. The inter-subject variability was apparent when examining the result of the listening test. The scores for the same transcription often ranged from 2-5 or 1-4 in score, with an average range of 2.8 steps. This acceptance of alternative pronunciations for the same word is more common for names than for other words. However there are some common words with multiple pronunciations, for example *paprika* [pʌ:prika , pʌprika] (ëpepperí) or *kex* [çeks , k'eks] (ëbiscuití).

Figure 23 shows the distribution of the numbers of subjects that thought that the transcription of the name was wrong. As can be seen none of the lexicon words had a majority of low (1-2) scores, while 10 of the 35 Twol generated transcriptions were regarded as wrong by the majority of the subjects. If the tendency for these 35 names holds for the rest in the test set of 1,000, only 1.7% of the transcriptions would be considered wrong by most people, compared the error rate of 6% mentioned earlier. If the same assumption is used for the rules about 20% of the transcriptions would be considered wrong. This is of course a hypothesis that has to be validated.

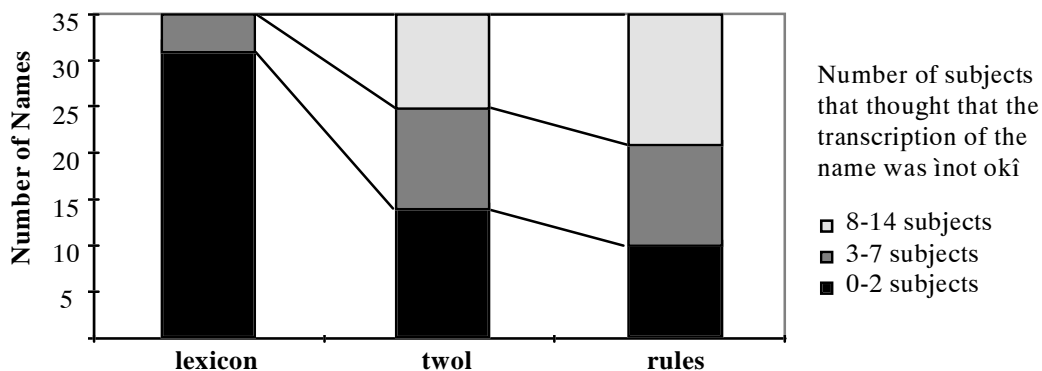


Figure 23. The number of names where the transcription were given the scores 1 or 2 by different number of subjects. None of the names in the lexicon were for example given these scores by a majority of the subjects.

The distribution of the scores for each name given by the subjects is quite complex, but if the names are divided into different groups the picture becomes clearer. The first group of names are the names where both of the automatic transcriptions were accepted by the majority of the subjects, see Figure 24.

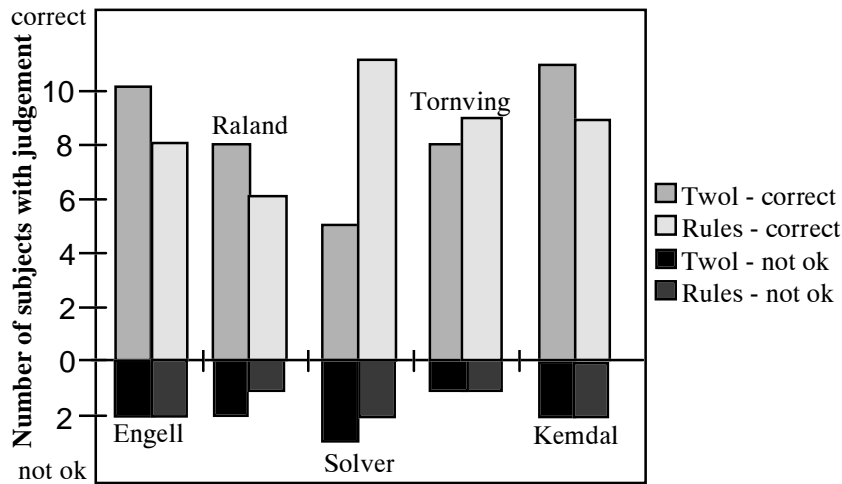


Figure 24. The distribution of the subjects' judgements of the 5 names where both of the automatic transcriptions were accepted by the majority of the subjects.

The names where all three transcriptions were acceptable for most of the subjects were of Swedish origin. Either the difference between the transcriptions was insignificant, for example as to whether there should be a morph boundary or not in the names *Raland* and *Tornving*; or the transcriptions were equally acceptable, like the pronunciations [u] and [ɔ] for the letter ööi in *Solver*.

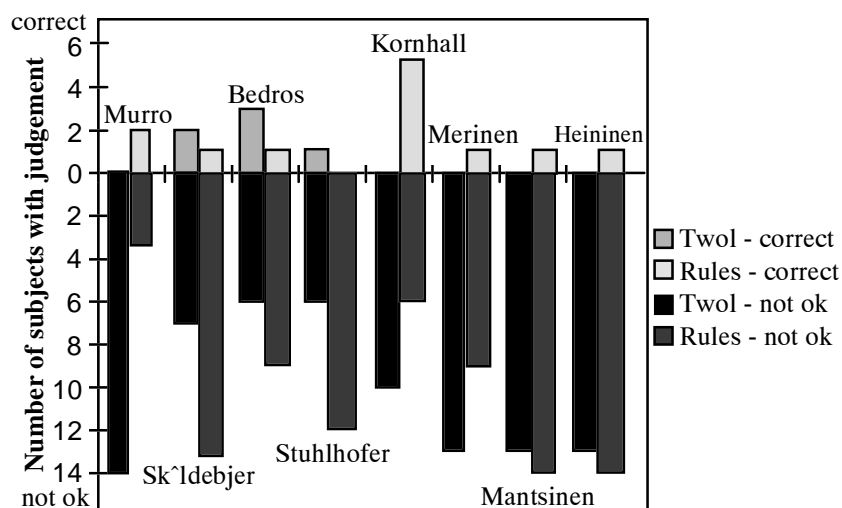


Figure 25. The distribution of the subjects judgement of the 8 names where both the automatic transcriptions were considered as wrong by the majority of the subjects.

The main reason for the Twol analyser to transcribe the names in Figure 25 wrong is that Swedish morphs are found in names of foreign origin. This often makes the morphological boundaries and primary stress misplaced. For example, the foreign name

Bedros have been analysed as *bed#ros*, giving the wrong transcription [b'e:d#r'u:s] instead of [b'e:drɔs] or [bedr'ɔs]. Finnish names like *Merinen* ending with *-en* have been incorrectly analysed according to the name pattern of Swedish names like *Nylen* [nyl'e:n]. This has resulted in the transcription [meri:n'e:n] instead of [m'e:rinən]. The rules are not designed for foreign names, which results in wrong transcriptions for these names. As mentioned earlier in this chapter, the names in this part of the corpus do not share name morphs with the low rank names that were used to develop the name specific rules.

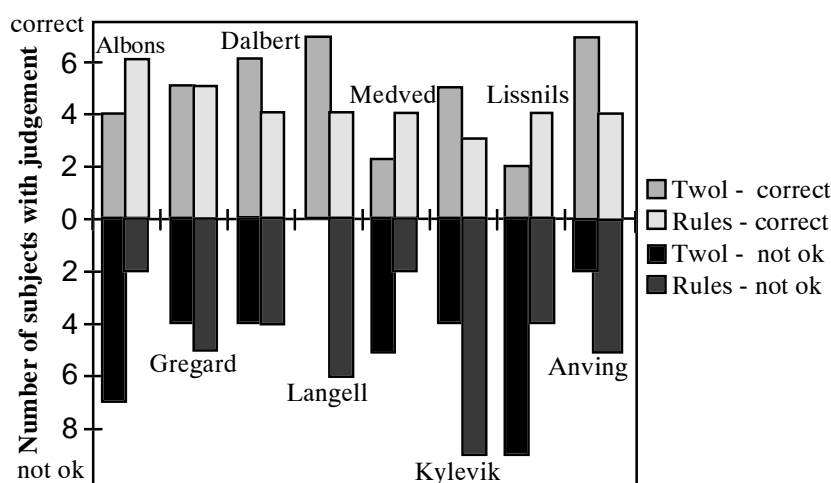


Figure 26. The distribution of the subjects' judgement of the eight names for which the subject had the most *ëvariedí* opinions of the same transcription.

For some names the subjects gave very different judgements on all transcriptions, see Figure 26. One reason for the difficulty in giving a single correct transcription for these names is that their origins are difficult to determine. They could either be foreign names looking Swedish, like *Medved*, or regional names from some part of Sweden, like *Lissnils* from the county of *Dalarna*. The reason for these names being hard to judge might be that they contain structures that differ from the structures found in the most common names. The most common names are of category I and II described earlier in the discussion on name analysis (chapter 4.1). The name *Kylevik* is of category II, but the difficulty lies in deciding whether the initial *ëkí* should be pronounced [k] or [ç].

The transcription errors can be divided into five groups: misplaced stress, wrong accent, morph errors, wrong phoneme and wrong vowel-length. According to Figure 27 the errors that got the lowest scores in the listening test are those with misplaced primary stress or with the wrong word accent.

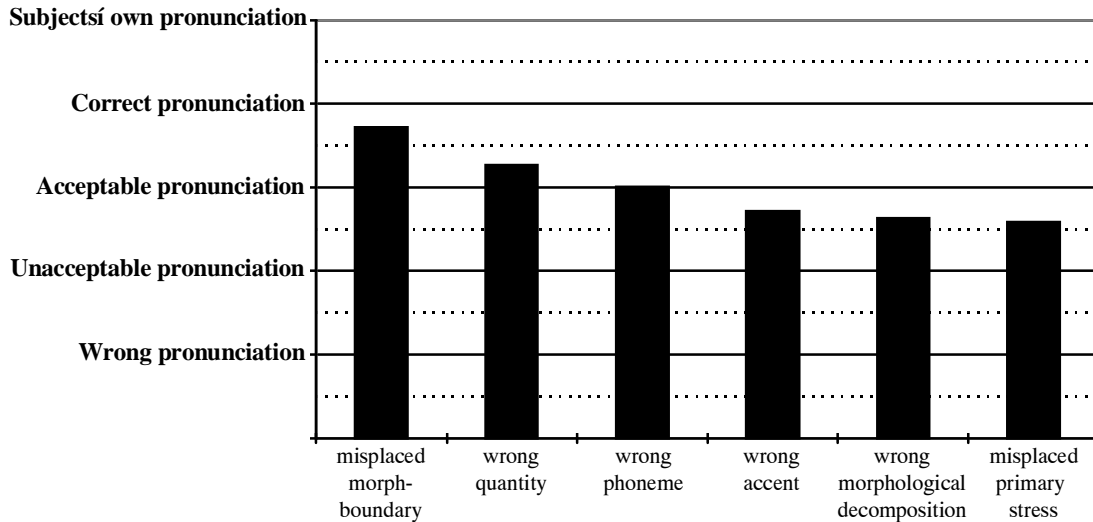


Figure 27. The average scores for different transcription error types .

Transcriptions with the wrong vowel length are often acceptable. This might be due to the fact that the realisation rules in our text-to-speech system will neutralise the duration difference in unstressed positions of +/-TENSE vowels, making the difference negligible. Another reason is that the length of the stressed vowel is dependent on the morphological analysis, for example where in consonant clusters the syllable boundary is positioned. The name *Dalbert* could either have the syllabification *dalb.ert* making the *äaí* short, or *dal.bert* making the *äaí* long. If the only error is a misplaced or missing morph boundary the difference in pronunciation is almost undetectable. If the position of the morph boundary influences other features it will make the acceptance of the transcription lower. Figure 28 shows that most names that were regarded as not correct in the evaluation are regarded as acceptable by a majority of the subjects.

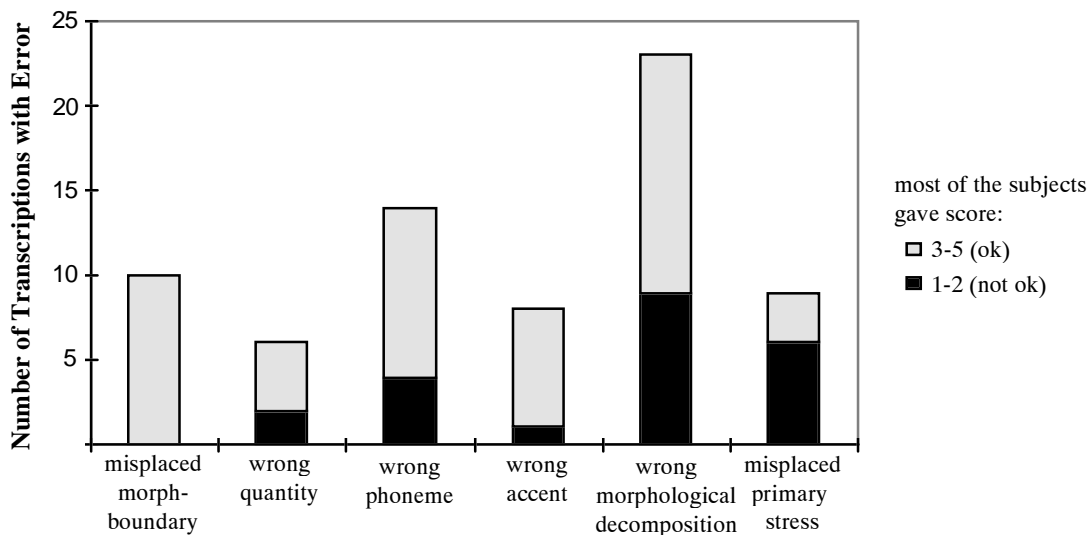


Figure 28. The distribution of the subjectsí judgements of names with different errors.

8.2. The Onomastica audit evaluation

Within the Onomastica project the qualities of all the transcriptions have been measured by an audit of independent auditors who were native speakers of the respective languages (ONOMASTICA, 1995). A total of 1000 names from each quality band were 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 5-7% included names not tagged as Swedish. Not all of these 5-7% incorrect transcriptions were wrong. Most of them were considered to be possible pronunciations, acceptable to a native speaker. They might not be equally acceptable, but some of them could very well have been produced in some dialect of Swedish. It was consequently considered safe to put these transcriptions in band II without checking. The result from the audit shows that this was appropriate, since no wrong transcriptions at all were found among the 1,000 names in the test sample from band II. The band I and II names had an error rate of only 0.3% and 0%. This exceptionally low error rate could be explained by the fact that only unacceptable transcriptions were marked as wrong in this audit.

9. Conclusions

The work described in this thesis has a bearing on speech synthesis and speech recognition of names. The main results of the work are:

1. The names in the Swedish telephone directory have been analysed and data about their structures have been collected.
2. A grapheme-to-phoneme conversion system for names have been developed.
3. The transcription system has been used to produce a pronunciation dictionary of almost 200,000 names, that can be used in various text-to-speech systems.

There are a number of factors that influence the pronunciation of a name, e.g., the origin of the name, the dialect of the speaker and the context in which it is produced. This work has used names that appeared in the telephone directory 1994 and they were transcribed according to the pronunciation of standard Swedish, i.e., the language of the author. New names are continuously introduced in Sweden either by the inventing of new names by Swedes, or by the introduction of foreign names by immigrants or from foreign, mainly American, movies. The spelling and pronunciation of foreign words are gradually changed according to the phonotactics of the native language. First name fashions change rather frequently. This causes parts of the pronunciation dictionary to be outdated after a short time. The Swedish name pronunciation system that has been developed will therefore be useful, since the rules are easy to adjust, according to some general pronunciation shift. New surnames that follow the structures of Swedish surnames, described in this thesis and implemented in the Twol approach, will be covered automatically.

The resulting dictionary and transcription system are already incorporated in our text-to-speech system. The grapheme-to-phoneme conversion in this system is done in three steps:

Lexical look-up	First the words are looked up in a domain-dependent dictionary, for example the Waxholm lexicon in the Waxholm project. Then the lexicon of the 107,379 most common Swedish non-name words is consulted. Finally a name-lexicon derived from the work described in this thesis is used.
Morphological decomposition	The words that were not found in any of the lexicons are first analysed by Twol using the morph lexicon for regular Swedish words. Then a second analysis is performed using the name-morph lexicon.
Language identification and to-sound conversion by rules	The remaining words are mostly of foreign origin. The origins of the words are obtained using the techniques letter- described previously. Then they are transcribed using the context dependent letter-to-sound rules of that language. Foreign phonemes that do not occur in Swedish are mapped to the closest Swedish phonemes.

A problem is how to find the names in a text (Wolinski, et al.,1995). They can be identified by examining the case of the words. If all words are in lower case, as often in electronic mails, some further methods of identification have to be introduced. A test has been performed that tried to identify Swedish names using trigraph statistics, in the

same way as language identification has been done. This approach identified 96% of the 87,000 Swedish names as names. The problem is that the names have the same patterns as other words. The difference is that some of these patterns are more common in names than in other words. Words containing these name morphs will consequently be identified as names by this approach. In our lexicon of 107,379 non-name words 36% were recognised as names. A safer approach is the one used in our text-to-speech system, where the name lexicons are used only for words not found in the other lexicons. In an ideal text-to-speech system all possible solutions are generated. Syntactic and semantic analysers are then used for disambiguation.

The Onomastica lexicon will also be used in various telephone services, since the Associated partners in the project are the telephone companies. A reverse directory enquiry service is already in operation in Italy as a public service handling several million calls per year. In Sweden a range of network services and facilities using the Swedish name database are being developed.

The Onomastica project will continue by including Eastern and Central European names - Czech, Estonian, Latvian, Polish, Romanian, Slovak, Slovenian, and Ukrainian. This will be done in a new project (COP-58) funded by the EC COPERNICUS Programme.

10. Acknowledgements

I would like to thank my supervisor Björn Granström, for all his support and the interest he showed for my work presented in this thesis. I especially want to thank Kjell Gustafson for sharing his knowledge, while proofreading this thesis. I am particularly grateful to Inger Karlsson and Nikko Ström for their patience with my sighs during parts of this project. I thank Rolf Carlson for inspiration and guidance. I thank Anders Lindström for the knowledge base he built during the previous name project. I would also like to thank my colleagues in the Onomastica project for interesting meetings.

Special thanks to my girlfriend Eva for her love and support, and to my family and friends.

I thank Rolf Carlson, Björn Granström, Kjell Gustafson, Jesper Håberg, Inger Karlsson and Eva Vanhainen for helpful advice concerning this text.

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

11. References

- Ainsworth, W A and Pell, B (1989). "Connectionist architectures for a text-to-speech system", Proceedings of Eurospeech 89 vol. 1 pp. 125-128.
- Allén, S and Wåhlin, S (1995). "Förnamnsboken: de 10 000 vanligaste förnamnen", Norstedt, Stockholm.
- Anderson, O and Dalsgaard, P (1995). "Multi-lingual testing of a self-learning approach to phonemic transcription of orthography", Proceedings of Eurospeech 95 vol. 4 pp. 1117-1120.
- Andersson, T (1979). "Svenska släktnamn i går, i dag - i morgon?", Nysvenska studier vol 59-60, pp 385-400.
- Andersson, T (1981). "Personnamn till begreppets avgränsning", NORNA80 (NORDic Name-symposium).
- Antwoth, E (1990). "PC-KIMMO A Two-level Processor for Morphological Analysis", Summer Institute of Linguistics, Dallas, Texas.
- Belhoula, K (1993). "Rule-based grapheme-to-phoneme conversion of names", Proceeding of Eurospeech 93 pp. 881-884.
- Basson, S Yashchin, D Kalyanswamy, A and Silverman, K (1993). "Comparing synthesisers for name and address provision: field trial results", Proceedings of Eurospeech 93 pp. 2165-2168.
- Basson, S Silverman, K Kalyanswamy, A Silverman, J and Yashchin, D (1993). "Synthesiser intelligibility in the context of a name-and-address information service", Proceedings of Eurospeech 93 pp 2169-2172.
- Blomqvist, M (1993). "Personnamnsboken", Oy Finn Lectura Ab, Loimaa.
- Byrd, R and Chodorow, M (1985). "Using an on-line dictionary to find rhyming words and pronunciations for unknown words", Proceedings of COLING 85 pp. 277-283.
- Carlson, R and Granström, B (1976). "A text-to-speech system based entirely on rules", Conference Rec. 1976 IEEE International-Conference on ASSP Philadelphia, PA.
- Carlson, R Elenius, K Granström, B and Hunnicutt, S (1986). "Phonetic properties of the basic vocabulary of five European languages: Implications for speech recognition", Proceedings ICASSP 86, Vol. 4, Tokyo, pp. 2763-2766.
- Carlson, R Granström, B and Lindström, A (1989). "Predicting Name Pronunciation for a Reverse Dictionary Service", Proceedings of Eurospeech 89 Vol. 1 pp. 113-116.
- Carlson, R Granström, B and Lindström, A (1990). "Automatic generation of name pronunciation for a reverse dictionary service", Report, Dept. of Speech Communication Music Acoustics, KTH.
- Carlson, R Granström, B and Hunnicutt, S (1991). "Multilingual text-to-speech development and applications", A W Ainsworth (Ed.), Advances in speech, hearing and language processing, JAI Press, London, UK.
- Church, K (1986). "Stress assignment in letter to sound rules for speech synthesis", Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing 4 pp. 2423-2426.
- Coker, C Church, K and Lieberman, M (1993). "Morphology and Rhyming: Two Powerful Alternatives to Letter-to-sound Rules for Speech Synthesis", Proceedings of the conference on Speech Synthesis, Aufrans, France 1990.
- Elert, C-C (1964). "Phonologic studies of quantity in Swedish", Almquist & Wiksell, Uppsala.
- Elert, C-C (1971). "Uttalsbeteckningar i svenska ordlistor, uppslags- och lexikon", Studier i dagens svenska, utgiven av Nämnden för svensk språkvetenskap 44.
- Eliasson, S (1979). "Expressiv gemining hos svenska hypokorismer och ellipsord", Nysvenska studier vol 59-60, pp 341-361.
- Esling, J (1990). "Computer Coding of the IPA: Supplementary Report", Journal of the International Phonetic Association 1990, 20:1.
- Garlén, C (1991). "Svenska ortnamn uttal och stavning", N&S, Stockholm.
- Gerritzen, D (1993). "Because it Sounds Nice : Choice of a First Name", Onomastica Research Colloquium, pp. 13-18.
- Golding, A and Rosenbloom, P (1993). "A comparison of Anapron with seven other name-pronunciation systems", report from University of Southern California.
- Gustafson, J (1994). "Onomastica - Creating a multi-lingual dictionary of European names", working papers 43, Lund University Department of Linguistics pp 66-70.

- Gustafson, J (1995). "Transcribing names with foreign origin in the Onomastica project", Proceedings of ICPHS 95 vol2 pp 318-321.
- Gustafson, J (1995). "Using Two-level morphology to transcribe Swedish names", Proceedings of Eurospeech 95 vol4 pp 2231-2234.
- Henrich, P (1989). "Language identification for the automatic Grapheme-to-Phoneme Conversion of Foreign Words in a German Text-to-Speech System", Proceedings of the European conference on speech Technology, 1989
- Hunnicut, S Meng, H Seneff, S and Zue, V (1993). "Reversible Letter-to-Sound Sound-to-Letter Generation Based on Parsing Word Morphology", Proceedings of Eurospeech 93 vol2 pp. 763-766.
- Karlsson, F (1990). "SWETOL: A Comprehensive Morphological Analyser For Swedish", manuscript, Department of General Linguistics.
- Koskeniemi, K (1983). "Two-Level Morphology: A general computational model for word form recognition and production", Department of General Linguistics, University of Helsinki.
- Kvillerud, R (1980). "F^rnamn i G^teborg- Namnskick f^r skolbarn f^dda 1958", G^teborg.
- Lawson, E (1984). "Personal Names: 100 Years of Social Science Contributions", Names, vol 32 No 1, pp 45-73.
- Lieberman, M and Church, K (1989) "Text Analysis and Word Pronunciation in Text-to-Speech Synthesis", Darpa workshop on speech and natural language 1989.
- Liljencrants, J (1969). "Speech Synthesiser Control by Smoothed Step Functions", QPSR-4, pp 43-50.
- Magnuson, T Granstr^m, B Carlson, R and Karlsson, F (1990). "Phonetic transcription of a Swedish morphological analyser", Proceedings of Fonetik-90, Phonum 1, Reports from the Department of Phonetics University of UmeÅ.
- Mengel A (1993). "Transcribing Names - Multiple Choice Task: Mistakes, Pitfalls and Escape Routes.", Onomastica Research Colloquium, pp. 5-9.
- Mill J S (1891). "A system of Logic", Longmans and Co. London.
- Modeer, I (1964). "Svenska Personnamn", Lund.
- NorÈen, A and NorÈen, E. (1907). "Svenska familjenamn vid b^rjan av 1900-talet", AB Ljus Stockholm .
- Otterbj^rk, R (1979). "Svenska f^rnamn", Esselte studium, Stockholm.
- Parfitt, S H and Sharman, A (1991). "A bi-directional model of English pronunciation" Proceedings of Eurospeech 91 vol. 2 pp. 801-804.
- Pirelli, V and Federici, S (1995). "On the pronunciation of unknown words by analogy in text-to-speech systems: an evaluation", Onomastica Research Colloquium.
- Rentzpopoulos, R and Kokkinakis, G (1991). "Phoneme to grapheme conversion using HMM", Proceedings of Eurospeech 91 vol. 2 pp. 797-800.
- Riley, M (1991). "A statistical model for generating pronunciation networks", Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing 1991.
- Schmidt, M Fitt, S Scott, C and Jack, M (1993). " Phonetic transcription standards for European names (ONOMASTICA)", Proceedings of Eurospeech 93 pp. 279-282.
- Sejnowski, T and Rosenberg, C (1987). "Parallel Networks that Learn to Pronounce English Text", Complex Systems, 1:145-168, 1987.
- Sigurd, B (1967). "SprÅkstruktur", Wahlstr^m & Widstrand, Stockholm.
- Social Security Administration (1985). "Report of distribution of surnames in the social security number file Sept 1, 1984", SSA Pub No. 42-004, April 1985.
- Spiegel, M (1985). "Pronouncing surnames automatically", Proceedings of AIOS 85 pp. 109-132.
- Spiegel, M Altom, M Macchi, M and Wallace K I (1988). "Using a mono-syllabic test corpus to evaluate the intelligibility of synthesized and natural speech", Proceedings of AVIOS 88.
- Spiegel, M Macchi, M and Gollhardt, K. (1989). "Synthesis of names by a demisyllable-based speech synthesiser", Proceedings of Eurospeech 93 vol 1 pp 117-120.
- Sullivan, K P H and Damper, R I (1991). "Speech synthesis by analogy: recent advances and results", Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing 1991.
- Statens offentliga utredningar (1964). "Svensk namnbok 1964", SOU 1964:14, Stockholm.
- StÅhle, C I (1979) "Kungsan, fikis och stegan", Nysvenska studier vol 59-60, pp 362-384.
- Svenska Akademien (1990). "Svenska Akademiens ordlista: SAOL:11", Norstedt, Stockholm.

- TegnÈr d.y. E (1882) "Om svenska familjenamn", Nordisk tidskrift.
- Trost, H (1993). "Morphology Workshop", Course notes from ELSNET Summer School on Prosody, London, 1993.
- Van Coile, B (1990). "Inductive Learning of Grapheme-to-Phoneme Rules", Proceedings of ICSLP 90 vol. 2 pp. 19.1.1-19.1.4.
- Van Coile, B Leys, S and Mortier, L (1992). "On the development of a name pronunciation system", Proceedings of ICASSP 92, pp 487-490.
- Viana, C Trancoso, I and Silva, M. (1995). "On the pronunciation of proper names and acronyms in European Portuguese", Onomastica Research Colloquium.
- Vitale, T (1991). "An algorithm for high accuracy name pronunciation by parametric speech synthesizer", Computational Linguistics, Vol. 17, No. 3, pp.257-76.
- Wennstedt, O (1995). "Namn i text", Licentiate thesis, (not yet published), University of UmeÅ.
- Wolinski, F Vichot, F and Dillet B (1995) ", Automatic Processing of Proper Names in Texts".
- WÅhlin, S (1977). "Variantstavningar av sl%oktnamn", Arbetsrapport sprÅkdata, G^teborgs universitet.
- Yvon, F (1993). "A Tidy Rule-Based Grapheme to Phoneme Transcriber for Onomastica", Onomastica Research Colloquium, pp. 13-18.
- Yvon, F (1994). "Self-learning techniques for grapheme-to-phoneme conversion", Onomastica Research Colloquium, pp. 25-38.
- The Onomastica Consortium (1995) "The Onomastica Interlanguage pronunciation lexicon", Proceedings of Eurospeech 95.
- "ONOMASTICA Multi-Language Pronunciation Dictionary of Proper names and Place Names", Technical and Financial Annex, Project No. LRE-61004.
- "ONOMASTICA Multi-Language Pronunciation Dictionary of Proper names and Place Names", Final Report, Project No. LRE-61004 (1995).

