UPPSALA UNIVERSITY

SCARRIE

Project ref. no. *LE3-4239*

Project title *SCARRIE Scandinavian Proof-reading Tools*

Deliverable number *DEL 4.1.3*

Deliverable title ***Specification of Phonemic Representation, Swedish***

Number of pages *12*

WP/Task responsible *Anna Sågvall Hein, Department of Linguistics, Uppsala University, Box 513, S-751 20 Uppsala, Sweden*
Email: anna@ling.uu.se

Author(s) *Leif-Jöran Olsson, ljo@stp.ling.uu.se*

EC Project Officer *Pierre-Paul Sondag*

Keywords *Grapheme to phoneme conversion, spell checking, spelling correction, CORRie*

Abstract *This report outlines the phonemic representation of Swedish as implemented for Swedish spelling correction with CORRie.*

EXECUTIVE SUMMARY

This task consists in the creation of a specification of phonemic representation for Swedish. This specification will be the basis for creating a grammar for the generation of a phonetic dictionary of Swedish. The grammar uses the formalism for grapheme to phoneme conversion available in the CORRie software package. The conversion grammar and the phonetic dictionary will be used as a part of a combined sound based and spelling based word level correction of Swedish.

Conclusion: The grammar produces accurate phonetic representations and suggestions for competence errors.

Some correction examples reproduced from chapter nine of the specification. The word following a > is the user input and the numbered words are possible corrections. All examples have the correct word as number one of the suggestions.

> jöra
1. göra [900]
2. gör [700]
3. för [600]
4. före [600]
5. förra [600]
6. fyra [600]

> grött
1. gått [600]

> kann
1. kan [700]
2. fanns [600]
3. hans [600]

> shu
1. sju [900]
2. så [600]

> jup
1. djup [900]

> tjåg
1. tjog [900]

> jystera
1. justera [800]

> diskution
1. diskussion [800]

> diskission
1. diskussion [800]

> konugn
1. konung [800]

Uppsala University
Department of Linguistics
Leif-Jöran Olsson
SCARRIE, 1998-01-21

Specification of Phonemic Representation, Swedish

DEL 4.3.1 of the SCARRIE Project

Abstract:
This report outlines the phonemic representation of Swedish as implemented for Swedish spelling correction with CORRie.

1. INTRODUCTION

This report outlines the phonemic representation of Swedish as implemented for Swedish spelling correction with the CORRie software package. The phonemic representation is used in the comparison of the input string and the representations of the words stored in the lexicon. This approach will handle both competence and performance errors.
The relations between phonemes and graphemes can be studied from two opposite angles:

i) how a grapheme is pronounced, **Graphophonematic Relations**, or
ii) how a phoneme is spelled in writing **Phonographematic Relations**.

In chapter two and three the relations between the phoneme and its spelling in Swedish is presented as summarised by Garlén, pages 157-162, [GAR88].
In chapter four there is an discussion of prosodic features. Some phonological processes are presented in chapter five. Chapter six deals with words with no relation between pronunciation and written spelling and in chapter seven there are some test words used during the development. Chapter nine shows some correction suggestions examples.

2. GRAPHOPHONEMATIC RELATIONS IN SWEDISH

The phonemes are represented by the IPA 1993 symbols [SIL] in this report. In the implementation the CELEX phonetic Alphabet, [CEL], has been used, with some modifications though, since the Swedish language has nine vowels and nineteen consonants in its alphabet. (Some of the vowel sounds were not represented in CELEX.)

Grapheme	**Represents**	**Examples**
a	/a/	bar, hal, barr, hall.
b	/b/	bo, tub, stubbe.
c	/s/	In front of ”e, i, y” in the same morpheme: ceder, citron, cykel.
cc	/ks/	In front of ”e, i y” in the same morpheme: accent, succé.
ch	/⌊/ or /X/ according to:	
		/⌊/ usually in words with French origin (where ”ch” normally represents /Σ+/): champinjon, chef, champagne, charm.
		/X/ usually in words borrowed from English and to some extent Spanish (where ”ch” normally represents /tΣ/): charter, cheddar, chips, chacha, chilen. In a few English words, e.g. cheviot, /⌊/ is the normal. A few words can have both /⌊/ and /X/: chans, choke, match.
	/k/	In the word och (and)
c	/k/	In other places, also in the combination ck: café, cowboy, cup, clown, crawl, flicka, suck.

d	/d/	dal, rad, lada, ladda.
dj	/j/	Morpheme initially: djup, djur, djävul.
	/dj/	In the word djonk.
e	/a/	In some words borrowed from French: engagemang, cendré
e	/e/	Usually when realised as ”long”: ed, se
	/E/	Sometimes (e.g. in the prefix er-, in names and foreign words): erkänna, erhålla; Per, Erling; intern, konsert.
e	/E/	Always short realisation in front of ”r”: berg.
e		When realised as ”short” in other places a lot of people use /e/ in some words and /×/ in other.
f	/f/	fel, tuff, soffa.
g	/j/ according to:	
		1. Morpheme initially or initially in stressed syllable in front of ”e, i, y, ä, ö”: ge, gissa, gynna, gäspa, göra. Exceptions for some foreign words in which /g/ is the norm, e.g. gejser, getto, gerilla, logik, region. In a few other foreign words /j/ alternates with /g/, e.g. agera, gigant, zigenare, or with /⌊/: gigolo, giraff, genialisk.
		2. After ”l” or ”r” in the same morpheme: alg, galge, helg, arg, berg, varg. Some exceptions are foremost proper names and foreign words: /lg/ in e.g. Helge, Algot, helgon. /rg/ in e.g. Borgå, embargo, largo.
g	/N/	Between a vowel and an ”n” in the same morpheme: agn, lugn, magnet, regn, vagn, hägn. Exceptions are: champagne,where gn(e) represents /nj/, and words like diagnos, prognos (in which ”gn” stands morpheme initially in the original language), which are pronounced /gn/.
g	/⌊/	In front of ”e, i” initially or in front of a stressed vowel in some foreign words, foremost of French origin: geni, genera, logi.
ge	/⌊/	Word final and in front of an other vowel grapheme: garage, prestige, sergeant.

gi	/⌊/	In front of other vowel grapheme: religiös.
g	/g/	In all other places: glad, gno, gris, gal, gol, gul, gå, egen, häger, mögel.
gj	/j/	Morpheme initially: gjorde, avgjord. Exception: Gjallarnornet which is pronounced /gj/.
h	/h/	hal, hur, behöver.
hj	/j/	hjort, hjärta.
i	/i/:	is, sil, ilska, sill.
j	/⌊/	In many foreign words of foremost French origin: jasmin, jargong, jour, projekt.
j	usually /j/	jaga, haj, höja.
		In addition to this ”j” occurs in various grapheme sequences which represents /⌊/, /X/ and /j/. See below.
k	/X/	Morpheme initially in front of ”e, i, y, ä, ö”: kela, kila, kyla, kära, köra. Exceptions are some foreign words and slang and children’s language words, which are pronounced /k/: kefir, kex, kille, kis, kissa, kisse.
ki	/X/	In kiosk (which has an alternate representation /k/).
kj	/X/	Morpheme initially in a few words: kjol, kjortel, kjusa (small valley).
k	/k/	In all other places: kal, kol, kula, kål, klar, krog. ”k” also occurs in the combinations ”sk” och ”skj” which represents /⌊/. See below.
l	usually /l/	lat, tal, mala, tall. In the combination ”rl”, ”l” is usually ”muted”: karl (man), värld.
lj	/j/	Morpheme initially: ljud, ljuga, ljus.
m	/m/	mat, tam, timme.
n	usually /n/	nål, lån, ana, hinna.
		In foreign words from French which ends in ”-nd, -ns, -nt” there are spellings with /-Nd, -Ns, -Nt/ as well as /-nd, -ns, -nt/,

		e.g. fond, chans, brons, patiens, genant, intressant. The pronunciation with /n/ is more frequent with younger speakers.
ng	/N/	Where ”n” and ”d” belongs to the same morpheme: säng, ånga, ring, sjunga.
o	/o/ eller /u/ according to:	
	/o/	Usually in the following stressed endings of foreign origin:
		-fon, e.g. grammofon, mikrofon, telefon **-for**, e.g. metafor, semafor **-ob**, e.g. mikrob **-of**, e.g. filosof **-og**, e.g. katalog, pedagog **-om**, e.g. agronom, astronom, idiom **-on,** e.g. elektron, metronom **-os**, e.g. narkos, neuros **-ot**, e.g. despot **-skop**, e.g. stroboskop, teleskop
	/u/	Usually in the following stressed endings of foreign origin:
		-od, e.g. metod, period -**onisk**, e.g. harmonisk, platonisk **-orisk**, e.g. notorisk **-orium,** e.g. sanatorium **-(t)ion,** e.g. union, passion, nation, station
	/o/	Usually in front of a ”v” in the same morpheme: dov, hov, lov, sova. Please observe that hov and lov have homographs which are pronounced with /u/.
		In ”short” realisation /o/ is the most usual, i.e. in front of a consonant combination or ”j” in the same morpheme: bock, torsk, holk, skoj.
		In ”long” realisation /u/ is the most usual representation (but not in front of ”v”): bo, fot, skog, tokig.
p	/p/	par, rep, apa, trappa
q	/k/	Only marginally in proper names and a few foreign words: Qvist.
r	/r/	rak, kar, mara, darra.
s	/s/	sak, tös, tuss, massa.

sc	/⌊/	In some foreign words: crescendo, fascist. Please observe the spelling with /s/ in scen, obscen.
sch	/⌊/	When ”s”, ”c” and ”h” belongs to the same morpheme: schakt, schism, dusch.
sh	/⌊/	In a few foreign words: shoppa, shunt.
shi	/⌊/	Between vowels in a few foreign words: fashionabel.
si	/⌊/	Between vowels in a few foreign words: division.
sj	/⌊/	Morpheme initially in a large amount of words, mostly domestic: sju, sjuk, sjö.
sk	/⌊/	Morpheme initially in front of ”e, i, y, ä, ö”: sked, skina, sky, skär, skön. Exceptions are some foreign words like skeptisk and skiss. ”sk” is also represented as /⌊/ in the words människa och marskalk.
skj	/⌊/	Morpheme initially in a few domstic words: skjuta, skjorta.
ssi	/⌊/	Between vowels in foreign words: mission
ssj	/⌊/	Between vowels within the morpheme: ryssja, vyssja.
stg	/⌊/	In the words gästgivare, västgöte and östgöte, and in their compounds and derivations.
sti	/⌊/	Between vowels in some foreign words: suggestion.
stj	/⌊/	Morpheme initially in the words: stjäla, stjälk, stjälpa, stjärna och stjärt.
t	/t/	ta, mat, matta.
ti	/⌊/	In some words ending in **-tion**, e.g. auktion, lektion, station, but /t⌊/ in others, e.g. motion, nation, portion.
tj	/X/	Morpheme initially: tjata, tjog, tjuta, tjära.
u	/← /	fura, sula, full, snurra.
u	/v/	Marginally in proper names and a few foreign words: Ouist
v	/v/	val, lav, näve.

w	/v/	Marginally in proper names and a few foreign words: watt, Wellander, wellpapp, weltervikt.
x	usually /ks/	ax, yxa
xi	/k⏐/	In the suffix –xion: reflexion.
xj	/k⏐/	In the proper name Växjö.
y	usually /y/	yr, hyra, syster.
y	/Ø/	In fyrtio and compounds and derivations of this word.
y	/j/	In a few foreign words: yoga, yoghurt.
z	/ts/	In foreign words with German or Italian origin: mezzosopran, Schweiz, pizza.
z	/s/	In all other cases: zink, zon, zoo, zoologi, zulu.
å	/o/	hål, låt, håll, mått.
ä	/E/	häl, lät, häll, mätt.
ö	/Ø/	föl, lös, föll, löss.

3. PHONOGRAHEMATIC RELATIONS IN SWEDISH

Phoneme	**Realised as**	**Examples**
/p/	”p, pp”	apa, pappa
/t/	”t, tt”	tåt, åtta
/k/	”k, ck”	kåk, rycka
/b/	”b, bb”	bar, tub, stubbe
/d/	”d, dd”	dåd, ladda
/g/	”g, gg”	gå, aga, agg
/m/	”m, mm”	tam, mamma
/n/	”n, nn”	nå, inne
/N/	”ng”	äng, sjunga (/◡ɣ/ betecknas ”gn”: ugn)

/f/	”f, ff”	får, soffa
/s/	”c, s, ss, z”	cykel, så, oss, zon
/X/	”k, ki, kj, tj, ch”	kär, kiosk, kjol, tjog, check
/ɭ/	”ch, che, g, ge, gi, ige, j, je, sc, sch, sh, shi, si, sj, sk, skj, ssi, ssj, stg, sti, stj, ti”	chef, apache, geni, bagage, religiös, beige, jour, damejeanne, crescendo, schack, shunt, fashionabel, division, sju, skön, skjorta, mission, ryssja, västgöte, suggestion, stjärna, station, (/kɭ/ betecknas med ”xi” i reflexion och med ”xj” i Växjö)
/h/	”h”	ha, hund
/v/	”v, vv, w, u”	väv, vovve, watt, Quist
/j/	”j, g, dj, gj, hj, lj, y”	jord, genast, djur, gjord, hjord, ljuga, yoga
/l/	”l, ll”	al, le, alla
/r/	”r, rr”	rå, orre
/i/	”i”	bi, sil, sill
/e/	”e”	se, sett
/E/	”ä, e”	säd, sätt, berg
/y/	”y”	fyra, fylla
/Ø/	”ö, y”	lös, löss, fyrtio (marginally)
/←/	”u”	ful, full
/u/	”o”	bo, ost
/o/	”o, å”	kol, hoppa, kål, åska
/a/	”a, e”	kal, katt, cendré

4. PROSODIC FEATURES

The only prosodic feature, which is systematically represented in the spelling of Swedish, is length. In the following sections some rules for length are described.

4.1 Rule for short stressed vowels

As a marker of short vowel quantity a following consonant is doubled:

i) between a stressed short vowel and another vowel, e.g. soppa, massa,
ii) between a short stressed vowel and morpheme boundary, e.g. hopp, vasst,
iii) between a short stressed vowel and ”l”, ”r” or ”n”+ vowel, e.g. kittla, uggla, vackla, äpple, teckna, vittna, öppna, bättre, offra, vackra.

In all other cases there is normally a single consonant after a short stressed vowel**,** e.g. tält, mast, flykt, orka.

The most important exceptions from this rule are as follows [GAR88] pages 162-164):

1. **”**m**”** is doubled – except in damm, lamm and ramm – only in front of a vowel, e.g. hemma (c.f. hem), samma (c.f. samla), ramma (c.f. ramla), rummet (c.f. rum).

2. **”**m**”** is not doubled in some other words, for instance amen, domare, döma, romare.

3. **”**n**”** is not doubled in the prefixes an- (e.g. anta) and in- (e.g. inse), the morpheme kun- (e.g. kungöra, kunskap) or in a few very frequent words: an, den, din, en, han, hon, honom, igen, in, kan , man, min, mun, män, sen (sedan), sin, sjön, snön, sven, ton (1000 kg), vän (noun), än.

4. **”**n**”** is not doubled in front of a morpheme boundary if the suffixes **-t** (neuter, participle or supine), -d (participle) or -de (preterite) follows: sant, känt, spänt, påmint; känd, spänd, påmind; kände, spände, påminde.

5. **”**j**”** is never doubled: haj, loj, vaja, kavaj, skojig.

6. **”**v**”** is almost never doubled. One exception is vovve (doggy).

7. In foreign words there are some exceptions to these rules. On the one hand *an expected doubling is not performed* e.g. artikel, cykel (bicycle), kapitel, plus, titel (even with a long vowel), and on the other one *unnecessary use of doubling where it is not needed to mark short vowel quantity* is found e.g. affär, grammatik och parallell.

4.2 General Rule for Pronunciation with Respect to Vowel Quantity

This is a general rule for pronunciation of vowels:

A vowel character in a stressed syllable is pronounced with long quantity if it is followed by at most one consonant within the morpheme.

The rule is valid for words like sko, bestå, stål, skåps (morpheme boundary between ”p” and ”s”) etc.

Exceptions:

i) Words with final ”m” and ”n” and some other frequent words (se section 4.1, rule 7 above),

ii) Many words with combinations of consonant characters within the morpheme, they still have a long vowel. This is true for:
a) Words with the sequences ”rd”, ”rl”, ”rn” and sometimes ”rt”, e.g. färd, hård, jord; Karl, sorl, pärla; barn, hörn, värn; fart, kart, vårta,
b) Words with combinations of one character and ”j”, ”l”, ”n” or ”r”, e.g. stödja, stävja; tavla, segla; vakna, vekna; segra, säkra.

iii) Vowel shortening sometimes occurs in front of the dental suffixes -t and –te**,** which is not seen in the spelling, e.g. högt, kokt, kokte, köpt, vitt.

5. PHONOLOGICAL PROCESSES

Some phonological processes working on consonants in Swedish:

1**.** Rule of Voicelessness

/d/ becomes [t] in e.g. köpte

2. Assimilation of /n/

/n/ may become a [m] and [N] in “min bil” and “min katt”, respectively.

3. Supradentalisation

/r/ transfers its place of articulation on the immediately following dentals and is thereby deleted.

6. NON DERIVABLE PRONUNCIATION AND SPELLING

Some frequent words have a complex relation between pronunciation and spelling. The pronunciation is not derivable from the spelling and the spelling is not derivable from the pronunciation. These words’ phonematic and graphematic forms have to be learned lexically:

Examples: att (infinitive mark), beige, brådska, dag, dagen, dagar, de, dem, den, det, dig, du, femtio, fredag, fyrtio, genre, gödsel, hade, henne, honom, huvud, jag, karl, konsert, kuvert, lade, ledsen, lördag, matsäck, med, mig, morgon, mycket, måndag, nio, nittio, någon, någonsin, någonstans, något, några, och , onsdag riksdag, riksdagsman, sade, sedan, sextio, sig, sjuttio, skjuts, skjutsa, staden, sådan, sådant, sådana, säga, söndag, tio, tisdag, tjugo, torsdag, trettio, vad (pronoun, adverb), vara (verb), var (verb), varit, värld, åttio, är.

7. TEST WORDS

These words have been used for testing during the development of the grapheme to phoneme rules for a spell checker of Swedish using the CORRie software package.

sju	sked	skicka
jama	jumper	joddla

göra	gädda	jämföra
balja	arg	torg
dvärg	svälja	gick
sjal	sjutton	själva
addition	pension	diskussion
garage	energi	giraff
tjock	kyrka	kälke
tjuv	kung	bank
dygn	många	omtänksam
blinka	gagna	välsigna
köksbord	häxa	razzia
analys	scen	precis
lova	tjog	konung
blått	skärm	penna
justera	ljus	djup

8. ERROR CATEGORIES

Here are the two topmost error categories:
1. competence errors: jöra (göra), skälv (själv), känst (tjänst), gågna (gångna), bållen (bollen), ursekta (ursäkta), kann (kan), dräckt (dräkt), nyj (ny), galmal (gammal)
2. performance errors: diskissoin (diskussion), förenign (förening)

See also [WED98] for the actual errors in our newspaper material.

9. EXAMPLE CORRECTIONS

The word following a > is the user input and the numbered words are possible corrections suggested by CORRie. All examples have the correct word as number one of the suggested corrections.

> jöra
1. göra [900]
2. gör [700]
3. för [600]
4. före [600]
5. förra [600]
6. fyra [600]

> grött
1. gått [600]

> kann
1. kan [700]
2. fanns [600]
3. hans [600]

> shu
1. sju [900]
2. så [600]

> jup
1. djup [900]

> tjåg
1. tjog [900]

> jystera
1. justera [800]

> diskution
1. diskussion [800]

> diskission
1. diskussion [800]

> konugn
1. konung [800]

REFERENCES

[CEL]: CELEX Phonetic Alphabet, Center for Lexical Information in Nijmegen, http://www.kun.nl/celex/index.html

[GAR88]: Garlén Claes: Svenskans fonologi, Studentlitteratur 1988

[HAM75]: Hammarberg B & Svensson B: Svenska som främmande språk- en lärobok, Sveriges Radios Förlag, 1975

[SIL]: Summer Institute of Linguistics, http://www.sil.org/ , for IPA93 fonts

[WED98]: Wedbjer Rambell et al.: Error Corpora Database, http://www.stp.ling.uu.se/~olgaw/errortyp/index.html