

ONOMASTICA

Telenor, 1999

This database contains original data from the Norwegian part of the ONOMASTICA project [1][2]. The data consist of pronunciations of first names, family names, company names, street names and place names.

The data also include a set of foreign names.

FILE FORMATS

=====

The text file format is UNIX and the character set used is ISO 8859-1 Latin 1.

CATALOG STRUCTURE

=====

catalog	Contains	File name(s)	#Entries

+ root			
	Copyright statement	copyright.txt	
	This file	README.txt	
+distrib-+			
+fnvn	first names	fnvnXXY.ono	239 393
+envn	Surnames	envnXXY.ono	82 416
+fnvd	double first names	fnvdXXY.ono	27 096
+gate	Street names	gateXXY.ono	88 161
+sted	Place names	stedXXY.ono	6 180
+bedr	Company names	bedrXXY.ono	100 179
+mix	Foreign names	xx_nw.ono	13 074

XX in file names is a running serial number. The first and surnames inventory is based on the Norwegian telephone directory. Higher serial number is associated with lower probability in the directory.

Y in file names indicate the name source and can take the following values for place names:

- x Postal place names
- k Place names from official cartography
- t Names used in the Telenor directory

The "mix" catalog contains files with "norwegian" pronunciation of

foreign names. xx in place file names is a nationality code for the words in the file:

uk = United Kingdom,
pt = portugal
se = sweden
fr = France
de = Germany
nl = Netherlands

DATA RECORDS

=====

The data are structured in records as follows:

Data(example):	Explanation

SOO:	: Record start
ENT:NO0000001	: National prefix (NO) and series number
LBO:Hansen	: Ortographic name
FQO:31391	: Number of occurrences in source material
NO0:"hAn\$s@n	: UPronunciation in SAMPA phonetic alphabety
NO1:	: Alternative pronunciation SAMPA
NO2:	: Alternative pronunciation SAMPA
QUO:1	: Quality level, se note 1)
WHO:MS,AF	: Transcriber ID(s)
ET0:NO	: Etymology (NO = Norwegian)
CT0:Surname	: Name class
EOO:	: Record end

1) QU0: 1:	Checked by phonetician, knowing the name
2:	Checked by phonetician, not knowing the name
3:	Automatic transcription only

PHONEME CODES

=====

Phoneme codes are in SAM Phonetic Alphabet (SAMPA) [3].

In addition to the phoneme codes defined for Norwegian, the following codes have been used to transcribe English names:

SAMPA	Example word

T	Tin (thin)
D	Dis (this)
dZ	dZin (gin)
aI	raIse (rice)
eI	reIn (rain)
OU	pOUk (poke)
Note: Non-standard: the correct SAMPA symbol is @U	

Syllabic consonants are followed by an /*/ (asterix). (This is not standard SAMPA format.)

n* example: Botn ""bOt\$n*

l* example: Hodsle ""HOd\$l*

Ambisyllabic consonants (consonants shared by neighbouring syllables) are indicated by including the consonant on both sides of the syllable boundary.

Examples:

Anne ""An\$n@

Inger ""iN\$N@r

Bjarne ""bjA:rn\$r@

[1] The Onomastica Consortium. The ONOMASTICA Interlanguage Pronunciation

Lexicon,

URL: <http://www.inesc-id.pt/pt/indicadores/Ficheiros/3225.pdf>

[2] Schmidt, M. / Fitt, S. / Scott, C. / Jack, Mervyn A. (1993):

"Phonetic transcription standards for european names

(ONOMASTICA)",

In EUROSPEECH'93, 279-282.

URL:

http://www.cstr.ed.ac.uk/downloads/publications/1993/Fitt_1993_a.ps

[3] Wells, J.C., 1997. 'SAMPA computer readable phonetic alphabet'.

In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997.

Handbook of Standards and Resources for Spoken Language Systems.

Berlin and New York: Mouton de Gruyter. Part IV, section B.

URL: <http://www.phon.ucl.ac.uk/home/sampa/norweg.htm>

ONOMASTICA

Telenor, 1999

Denne databasen inneholder originaldata fra ONOMASTICA-prosjektet. Tekstfilene er lagret på Unix-format og tegnsettet som benyttes er ISO8859-1 Latin 1.

Her finnes uttale av fornavn, etternavn, gatenavn, stedsnavn og bedriftsnavn fra telefonkatalogen. Man har benyttet SAMPA fonetisk kode.

For en nærmere beskrivelse av materialet se rapporten "Norsk deltakelse i ONOMASTICA-prosjektet" (SINTEF DELAB rapport STF40 F95017).

KATALOGSTRUKTUR

=====

Katalog	Innhold	filnavn

+		
	Copyright statement	copyright.txt
	Denne filen	LESMEG
+distrib++		
+fnvn	fornavn	fnvnXXY.ono
+envn	etternavn	envnXXY.ono
+fnvd	doble fornavn	fnvdXXY.ono
+gate	gatenavnfiler	gateXXY.ono
+sted	stedsnavn	stedXXY.ono
+bedr	bedriftsnavn	bedrXXY.ono
+mix	Utenlandske navn	xx_nw.ono

I filnavnene er XX et løpende serienummer. Dess lavere nummer dess høyere antall forekomster i telefonkatalogen. Alle for- etter og bedriftsnavn kommer fra Telenor (telefonkatalogen).

Stedsnavn er hentet dels fra telefonkatalogen, kartverket og Telenors opplysningstjeneste.

Y angir kilde og kan ha følgende verdier:

- t Poststeder i telefonkatalogen
- k Tettsteder fra Statens kartverk
- x Stedsnavn som brukes av opplysningstjenesten i Telenor

Dersom Y ikke er inkludert i filnavnet er ikke uttalen i filene kontrollert av fonetiker. Dvs at uttalen kun er generert av et tekst-til-fonemprogram. Disse filene har derfor relativt lav kvalitet. Ingen av bedriftsnavnene har kontrollert uttale.

"mix"-katalogen inneholder filer med "norsk" uttale av utvalgte utenlandske navn.
filnavnene har følgende konvensjon

xx_nw.ono der xx er nasjonalitetsbokstaver (uk,pt,se,fr,de,nl)

DATAPOSTER:

=====

Hver fil har poster med følgende struktur:

Nøkkel:Data(eksempel) : Forklaring

SOO: : Start på datapost
ENT:NO0000001 : Nasjnal prefiks (NO) + serienr.
LBO:Hansen : Navn
FQO:31391 : Antall forekomster
NO0:"hAn\$ss@n : Uttale i SAMPA fonetisk alfabet
NO1: : Ev. alternativ uttale i SAMPA
NO2: : Ev. alternativ uttale i SAMPA
QUO:1 : Kvalitetsnivå se 1)
WH0:MS,AF : Transcribør ID-er
ET0:NO : Etymologi (NO: Norwegian)
CT0:Surname : Navneklasse
EOO: : Datapost slutt

--

1) QU0: 1: kontrollert av fonetiter, som kjenner navnet
2: kontrollert av fonetiker, men navnet er ukjent
3: kun automatic transripsjon

FONEMKODER

=====

Fonemkoden for navnene er angitt i SAM Phonetic Alphabet (SAMPA).
Dokumentasjon av SAMPA er å finne på internetadressen

<http://www.phon.ucl.ac.uk/home/sampa/norweg.htm>

I tillegg til fonemkodene som er definert for norsk, er det benyttet følgende koder fra det engelske foneminventaret for å transkribere engelske navn.

SAMPA	Eksempelord
T	thin
D	This
dZ	Gin
aI	rise
eI	raise
OU	nose (riktig SAMPA-kode er @U)

Stavelsesbærende konsonanter markeres

n* eksempel: Botn ""bOt\$n*
l* eksempel: Hoddle ""HOd\$l*

Deling av konsonanter mellom to stavelser er angitt ved å føre opp konsonanten på hver side av stavelsesgrensen

Eksempler:

Anne ""An\$n@
Inger ""iN\$N@r

Bjarne ""bjA:rn\$rn@

***** Per Olav Heggveit 13.05.96 *****
*****Jon Emil Natvig 15.3.1999*****