A Norwegian letter-to-sound engine with Danish as a catalyst

Peter Juel Henrichsen
Center for Computational Modelling of Language (CMOL)
Copenhagen Business School
Dalgas Have 15, DK-2000 Copenhagen F, Denmark
pjuel@id.cbs.dk

Abstract

Danish phonetics and Norwegian phone-tics are not all that different. This fact is exploited in an ongoing project establishing a phonetic transcription algorithm for (East-) Norwegian. Using methods known from machine learning we exploit a publicly available phonetic database for Danish (based on the Danish PAROLE corpus) arriving at a cost-competitive phonetic database for Norwegian. While the ultimate goal of this enterprise is a low-budget complete phonetic transcription of the NoTa corpus of Norwegian spontaneous speech, this paper presents the subparts related to Danish phonetics.

1 Introduction

Politically, Norwegian and Danish[1] are treated as distinct languages as a matter of course. From a linguistic point of view the distinctness is more dubious: it is not difficult to suggest a pair of Norwegian dialects differing more from each other (at least phonetically) than do the vernaculars of Copenhagen and Oslo[2]. Similarly, in a recent study based on statistical methods, Danish and Swedish spoken languages were shown to be in many respects better described as mutual dialects than distinct tongues (Henrichsen et al 2005). Recycling of linguistic resources across Scandianvian boundaries thus seems to be a natural idea.

In this short paper we present our Danish-to-Norwegian phonetic mapping algorithm based largely on publicly available resources. A dedicated homepage is found via a link in the author's homepage,
www.id.cbs.dk/~pjuel

The paper is sectioned as follows. In section 2 we introduce the Danish resources that have played a role in the present project. In section 3, the algorithm is presented, while section 4 presents some results. Finally section 5 puts the letter-to-sound enterprise in a wider perspective.

2 The Danish PAROLE project

The Danish PAROLE corpus was established in the early nineties as a part of the pan-European PAROLE project representing all the official EU languages. In each participating country, a local project group was appointed and commissioned to establishing a (by that time's standards) large text corpus. A substantial subpart of this corpus (at least 250k words) was then manually annotated for part-of-speech using a common annotation format allowing for easy transfer of PoS information across language boundaries.

Annotation work on the Danish PAROLE corpus has been continued at CMOL. Today PAROLE is a poly-dimensional corpus structure comprising these annotation dimensions (in various stages of completion):

- Tree structure (dependency based)
- Rhetorical structure

[1] In this paper, 'Norwegian' and 'Danish' refer to the East-Norwegian dialect spoken in and around Oslo, and the lect termed “advanced standard Copenhagen”, respectively – unless otherwise stated.

[2] In Denmark, the former dialectal diversity is now largely lost, probably due to the denser and more evenly distributed Danish population.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 305–309

- English translation
- Russian translation
- Tamil translation
- Phonetic annotation (*)
- Prosodic annotation (*)
- Sound track (by one male speaker), including:
  - *Fundamental frequency measurements* (*)
  - *Intensity measurements* (*)

Of these, the dimensions marked with * have played a role in the project reported. Contact the author for information on access to the Danish PAROLE corpus including the various annotation dimensions.

3 The transfer component

The transfer component includes several strategies. Those involving Danish are shown in fig. 1.

Figure 1. Norwegian letter-to-sound mapper

The topmost box represents the Norwegian orthographic input. One of two paths may be tried, in this order: (i) via NO2DO, DO2DP and DP2NP to the end state (the phonetic form), or (ii)via NO2NP. The modules are introduced below. As will be explained, DO2DP as well as DP2NP are informed by the phonetic forms inventory called *DanPO*, the Danish phonetic dictionary developed at CMOL as a part of the long-term PAROLE annotation project.

3.1 **Orthographic preliminaries**

It is conventional knowledge that many Norwegian orthographic forms have identical Danish equivalents; examples are “notat”, “notere”, “ignorerer”, and “ignorant”, just to mention a few. Other forms have near-equivalents, such as “infisere” (Danish *inficere*), “notasjon” (*notation*), “ignorerte” (*ignorerede*) and “ignorantene” (*ignoranterne*) differing only superficially. Only a very small fraction of Norwegian stems are completely absent in Danish, as can be verified in any ordinary word frequency list.

Inspired by these facts, a Norwegian-Danish orthography-to-orthography mapping (NO2DO) was established in a preparatory stage. In the text box below some of the most productive NO2DO rules are presented as (Perl-style) regular expressions.

NO2DO(Nor. orthograpy to Dan. ortography)

Orthographic surface rules

General equivalents
s/kj/k/, s/kt/gt/, s/øy/øj/, s/au[dt]/ød/, ...

Morphological equivalents
s/sak/sag/, s/skap/skab/, s/sjon/tion/, ...

Rules informed by PoS

Verbs: s/a$/e/, s/ert$/eret/, ...
Nouns: s/ene$/erne/, s/a$/en/, ...
Adjectives: s/aktig$/agtig$/, s/ert$/eret/, ...

Legend
s/*x*/*y*/ = substitute *x* by *y*
...$/ = matches string-final position only
[*xy*] = matches one instance of either *x* or *y*

Table 1

Using NO2DO conversion, the recognition rate of Norwegian-Danish equivalent word forms (based on a reference list of 1000 hand-picked equivalences) rose from an initial 75% precision and even lower recall, to more than 95% for both precision and recall[3].

3.2 **Danish phones to Norwegian phones**

In order to exploit our Danish source as far as possible, we based the mapping algorithm on a phonetic alphabet relating as closely as possible to the Danish PAROLE phonetics annotation (cf. author's homepage). This means that we had to deviate slightly from the de facto standard of SAMPA[4], the most substantial difference being that the retroflexes are not fully instantiated in the adopted version. Discussions with several Norwegian phoneticians have made us aware that the

[3]The actual figures should be taken with a grain of salt: the procedures of hand-picking and manual rule-adjustment may sometimes lead to artificially boosted results; still a considerably improved recognition rate is beyond doubt.

[4]Cf. www.phon.ucl.ac.uk/home/sampa

patterns of retroflexation are not completely consistent among East-Norwegian speakers (or even among Oslovians). The full definition of the Norwegian sound alphabet can be consulted at the Dphon2Nphon homepage (via link above).

In the following we demonstrate our treatments of a few of the most important systematic differences between Danish and Norwegian phonetic realisation of common underlying phonological forms.

In Norwegian, the stops /p/ /t/ /k/ are always realised as [p] [t] [k]; Danish /p/ /t/ /k/, in constrast, are only maintained in syllable-initial positions before full wowels; in other positions they reduce to [b] [d] [g], as marked in **bold** in fig. 2.

	Norwegian	*Danish*
"titte"	[t"it0]	[t`i**d**0]
"statsmakter"	[st'A:tsmAkt0r]	-
"statsmagter"	-	[s**d**`{:?**d**smA**gd**C]
"straffbart"	[str"AfbA:rt]	-
"strafbart"	-	[s**d**r`AfbA:?**d**]

For Norwegian: ['] = accent I, ["] = accent II.

Figure 2. Danish-Norwegian equivalent forms

The Danish *stød* (sometimes described as a quick glottal contraction or even a glottal stop, but actually better described as instance of 'creaky voice') has no articulatory equivalent in Norwegian. By and large, Danish stød corresponds to Norwegian accent I (as exemplified in fig. 2). The general correlation pattern is however overruled by two facts: (i) in Norwegian, the distinction between accent I and II is only relevant in polysyllabic words; Danish has no similar restriction for stød (e.g. "mand"/"man" [m`an?]/[m`an], "Hans"/"hans" [h`an?s]/[h`ans]; (ii) Danish stød only occurs in syllables with either a long vowel or a short vowel followed by a sonorant consonant ("tænger" [t`EN?C], "koen" [k2o:?0n], but not in e.g. "takker" and "hoppen"). No similar restriction applies for the Norwegian accents.

These and numerous other productive subsitutions rules (165 in all) have been identified by automatic and semi-automatic methods and incorporated in the DP2NP mapping algorithm.

DP2NP(Dan. phonetics to Nor. phonetics)

s/`*D*?/'*NVow*/ (select accent I)
s/d/t/ for "t" in Danish orthography
s/D/d/ for "d" or "t" in Dan. orthography
s/b/p/ for "p" in Dan. orthography
s/g/k/ for "k" in Dan. orthography
s/v/N/ for "gn"

Legend
`*DVow*? = matches any stressed vowel with stød
[D] = Danish /d/ without stop

Table 2

By way of example, Norwegian "adoptant" transcribes to [AdOpt'Ant] in these steps:

"adoptant"
*via***NO2DO**
"adoptant"
via **DO2DP**

```
[adCbt`an?d]
   ||  || |
[AdOpt'An t]
```

*via***DP2NP**

Observe that (in this case) the proper accent I is selected as signalled by the Danish stød.

3.3 **Norwegian orthography-to-phonetics**

Even if most Norwegian word forms by far have Danish equivalents, not all of them are within reach of automatically derived (i.e. machine learned) rules. A rule discovering the equivalence of, say, Norwegian "tvil" and Danish "tvivl" (*doubt*) might also – and erroneously – postulate the equivalence of "sal" (*hall*) and "savl" (*saliver*). Foreign words tend to maintain their original spelling in Danish ("bassin", "orange"), but not in Norwegian ("basseng", "oransje"). Finally, some Norwegian lexemes do not have equivalents in contemporary Danish, such as "kanskje" and "slik".

In all such cases, the Danish-to-Norwegian transliteration regime clearly does not suffice. Thus a third mapper had do be designed for the cases where no Danish equivalents can be identified, the Norwegian letter-to-sound module (NO2NP). This module cannot boast innovation, neither in function nor implementation, so we shall not care to present the details here. It follows principles and details from the literature (e.g. Sahajpal (05), Andersen (96), Black (91)). Our results are

probably comparable to those reported in the literature, as far as can be judged from the sparse information given; but we wish to stress the fact that we only included the NO2NP for our transformation algorithm to be complete. Details and references will be presented in Henrichsen (2007).

3.4 **Compounding**

Of course, rules for analysing lexically unrecognized words using rules of compounding have also been implemented. However, as such rules are not special to the current project, we do not present the details here. Suffice it to say that the Danish glue elements (fuger) “e” and “s” are matched by virtually similar Norwegian equivalents “e” and “s”. The Norwegian s-fuge usually does not alter the accent of the first component while the e-fuge usually produces accent II. Likewise, fuge-e induces loss of stød in Danish. Otherwise, the implications of compounding are very similar to the Danish rules (e.g. main stress retained by the first compound element only).

4 Some results

We selected 50,000 word forms randomly from the Oslo corpus of mixed text genres (cf. www.tekstlab.uio.no/norsk/bokmaal). Of these, 23,939 were processable by the NO2DO-DO2DP-DP2NP strategy (i.e. recognized in DanPO after NO2DO treatment). We then scored the results using three success definitions: *(i)* exact match with the phonetic form found in a standard Norwegian phonetic dictionary; *(ii)* a single phonetic conflict permitted (e.g. [e] mistaken for [0], or long-instance for short-instance of the same vowel), *(iii)* exact match when ignoring type-of-accent. Table 3 presents our results.

Word length	*(i)-correct*	*(ii)-correct*	*(iii)-correct*
2-5	78%	89%	84%
6-10	68%	79%	75%
11-15	59%	68%	66%

*Table*3

As seen, we get a good estimate of the Norwegian phonetic form of short words (up to ninety percent accuracy, relaxing the success criterion a little). Also, most errors made are not very disturbing (many are recurrent errors in transcriptions made by humans too), e.g. the confusion of [e]/[0], [o]/[u], accent I/II. As the large majority of words given the NO2DO-DO2DP-DP2NP treatment are shorter than 10 letters after compound decomposition, our preliminary conclusion is optimistic. The low-budget Norwegian phonetic annotation engine may be within reach.

5 Discussion

Phonetic translating between Scandinavian (or other cognate) languages is interesting in its own right, providing a constructive and directly testable way of pursuing 'micro-typology' and language history. To this comes the practical usefulness. The sub-project reported here together with its results constitute the first steps in an ongoing larger scale annotation project which as its primary goal has the phonetic annotation of the newly published NoTa corpus of about 900,000 words (cf. http://www.tekstlab.uio.no/nota/). The annotation task must be carried out on a very tight (almost non-existing) budget, so the recycling of readily avaliable resources for Danish including phonetics and even prosodic markup is – for several reasons – an attractive strategy.

The new NoTa data-tier is of course of lesser value to the linguist than real descriptive phonetics would be. However, for certain purposes, lexical phonetics may well suffice. Say you want to investigate the cross-speaker realisation of a particular phoneme or phoneme-combination that cannot be reliably traced by its orthographic image – then a lexical-phonetic search dimension may be just what you need.

Acknowledgements

Thanks to Janne Bondi Johannessen, Torbjørn Nordgård, and many others for loads of information and good sparring. Without them being Norwegian this project would have been impossible.

References

Andersen, Ove; Roland Kuhn; Ariane Lazaridès; Paul Dalsgaard; Jürgen Haas; Elmar Nöth (1996) *Comparison of Two*

Tree-Structured Approaches for Grapheme-to-Phoneme Conversion; ICSLP 1996 (cd-rom), 4pp

Black, A; Joke van de Plassche, Briony Williams (1991) *Analysis of Unknown Words through Morphological Decomposition*; proceed. of 5th EACL, 101-106

Grønnum, Nina (1998) *Fonetik og Fonologi - Almen og Dansk*; Copenh.: Akademisk Forlag

Henrichsen, Peter Juel; Jens Allwood (2005) *Swedish and Danish, Spoken and Written Language - a statistical comparison*; J. of Corpus Ling. 17/3:2005

Henrichsen, Peter Juel (2007) *NoTa – nu med lydskrift;* in book on NoTa (Oslo Univ., ed. by Janne Bondi Johannessen), *in prep*.

Nordgård, Torbjørn (2000) *NorKompLeks. A Norwegian Computational Lexicon*; COMLEX-2000, Patras, Hellas

Sahajpal, Anurag; Terje Kristensen (2005) *Transcription of Text by Incremental Support Vector Machine*; proceed. of Norsk Informatikkonferanse 2005, 79-87

Skadhauge, Peter Rossen; Peter Juel Henrichsen (2005) *DanPO - a transcription-based dictionary for Danish speech technology*; *NODALIDA-2005*, Joensuu, Finland