DanPO – a transcription-based dictionary for Danish speech technology

Peter Rossen Skadhauge and **Peter Juel Henrichsen**
CMOL
Department of Computational Linguistics
Copenhagen Business School
Denmark
prs@id.cbs.dk and pjuel@id.cbs.dk

Abstract

We present a new strategy for the creation of phonetic lexicons. As we argue, lexical resources for speech technology integration should be informed by transcriptions of spontaneous speech. We illustrate our strategy with examples from the dictionary DanPO (Danish Phonetic-Orthographic Dictionary) which is developed at the Center for Computational Modelling of Language (CMOL). For reference corpus we used DanPASS consisting of 57 recordings of task-oriented monologs, transcribed by professional and MA-level phoneticians using the Danish SAMPA phonetic alphabet. From the transcriptions, dictionaries and concordances were compiled, and these resources were merged with the (prescriptive) phonetic renderings of a standard Danish word dictionary of 87,000 lemmata. As an effect of the "transcription informed" strategy, DanPO is expected to significantly improve the success rate of automatic speech recognizers, as well as the naturalness of artificial voices. Furthermore, we devise an experimental strategy in order to evaluate the dictionary and further improve later versions.

1 Introduction

In this paper we present a novel approach to data-driven lexicography exploiting transcriptions of spontaneous speech as raw material. Our methods are being developed and tested in connection with our work on the Danish speech technological dictionary DanPO (Skadhauge and Henrichsen, 2005).

Formally speaking, DanPO (Danish Phonetic-Orthographic Dictionary) is an add-on to the general Danish language technological dictionary STO ("Sprogteknologisk Ordbase", "Lexical Database of Danish for Language Technology Applications", cf. (Braasch, 2003)). STO contains about 87,000 lemmata annotated with full inflectional and compound morphology as well as syntactic information (e.g. verb complement frames and semantic features). The STO dictionary was initiated by the Danish Ministry of Research. It was developed by researchers from a number of Danish universities, coordinated by the Center for Language Technology.

As a supplement to the STO dictionary, CMOL (Center for Computational Modelling of Language) is developing a phonetic computational dictionary DanPO. DanPO is distinct from a traditional paper-based phonetic dictionary in several ways:

- DanPO is generative, in the sense that any word or word form (including compounds) recognized in STO can be phonetically transcribed using the sound rules and inflectional information in DanPO.
- DanPO can be rendered as simple text files for easy embedding within speech technological products (e.g. artificial speech or automatic speech recognition).
- Our policy is 'open source', meaning that any party, be it private, institutional or commercial, will be allowed access to DanPO on friendly conditions (for a nominal fee).
- The phonetic transcriptions in DanPO are informed by actual transcriptions of spontaneous speech.

The DanPO project group consisted of two professional (computational) linguists and five

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 186–192
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

student transcribers. It was functioning for about two years. The first release of DanPO is scheduled for October 2005. Several industrial partners are taking part in the development of DanPO as external evaluators.

The formal properties of DanPO are presented in section 2 below: its internal structure and its embedding in STO. In section 3, we motivate the "transcription informed" strategy that we have adopted, while section 4 contains a short status report for project DanPO. Section 5 describes development strategies for later versions of the dictionary. Finally, in section 6, we draw some (preliminary) conclusions.

2 Formal properties

The key design choices of DanPO are the following:

- Any word form recognizable or producible by STO must be recognizable or producible with the information in DanPO.
- The internal structure of DanPO must mirror the internal structure of corresponding STO parts as closely as possible.
- The phonetics for the majority of lemmata must be generated from existing resources, thus minimizing the need for phonetic hand-coding.

The main obstacle for obtaining parallelism between orthography and phonology is the mismatch between orthographic and phonological inflectional paradigms. Each lemma in STO is associated with at least one of 675 inflectional paradigms. The majority of these paradigms account for irregular and semi-regular inflection of minor categories of words. A few paradigms account for the inflection of all regular lemmata.

The correspondence between Danish phonetics and orthography is complex and irregular. Several distinctive phonetic features, such as stress, vowel length, and "stød" (a quick glottal contraction) are not fully predictable from orthography. Thus, we expect to have to account for a certain amount of mismatch between the structures of orthographic inflectional morphology and phonetic inflectional morphology.

The paradigm `ORP0028` in Fig. 1 accounts for a class of common nouns, exemplified by "dag" (Eng.: "day"). The STO version of the paradigm covers 3485 lemmata:

The two exemplified subparadigms differ with respect to "stød" expressed by exclamation marks `[!]`. Nouns like "dag" `[dz:]` which bear "stød" in singular forms, fall into the paradigm of `ORP0028.1`, whereas nouns like "hest" `[hEsd]`, (Eng.: "horse") which lack "stød", belongs to the paradigm of `ORP0028.2`.

In total, about 20 subparadigms express similar systematic phonetic differences between forms of the orthographic paradigm of `ORP0028`.

On the other hand, many of the phonetic subparadigms are similar across categories of orthographic paradigms. That holds for the orthographic paradigms which double final consonants, e.g. ("slot", "slottet", "slotte", "slottene", ...) and ("stop", "stoppet", "stoppe", "stoppene", ...), where the orthographic consonant duplication has no phonetic counterpart.

The phonetic notation of DanPO is derived from The SAMPA computer readable phonetic alphabet (Wells, 1997). The notation of suffixation corresponds to the "search-and-replace" mechanism for PERL (Wall et al., 2000) regular expressions in the following sense: Every suffix consists of a search string and a replacement string. Thus the suffixation can handle phenomena related to vowel length and "stød" of the vowel in the final syllable of the stem.

As an example, the imperative is the only form of the verb "tegne" `[tAJn0]` (Eng.: "draw") which contains a "stød" on the vowel. The pair of search string and replacement string defines the phonetic properties of the relevant subparadigm of the orthographic paradigm. In this case, the search string contains two subpatterns; the first pattern being a stem whose last vowel is a diphthong, the second being an optional syllable-final consonant. The replacement string returns the strings matched by the first and second subpattern with a "stød" `[!]` in between.

Search string:
`(.+[\#V][\#S])([\#C]?)0`

Replacement string:
`\1!\2`

`#V` is the set of vowels, `#S` is the set of sibilants, and finally `#C` is the set of consonants.

The subpatterns (in parentheses) match `[tAJ]` and `[n]` respectively, which are reproduced by the "duplication" strings `[\1]` and `[\2]`. This

Figure 1: Sample paradigms of STO and DanPO

A STO paradigm (`ORP0028`**)**

```
ORP0028:dag:NOUN:COMMON:SINGULAR:INDEFINITE:NOMINATIVE::
ORP0028:dag:NOUN:COMMON:PLURAL  :INDEFINITE:GENITIVE  ::es
ORP0028:dag:NOUN:COMMON:SINGULAR:DEFINITE  :GENITIVE  ::ens
ORP0028:dag:NOUN:COMMON:PLURAL  :INDEFINITE:NOMINATIVE::e
ORP0028:dag:NOUN:COMMON:PLURAL  :DEFINITE  :NOMINATIVE::ene
ORP0028:dag:NOUN:COMMON:PLURAL  :DEFINITE  :GENITIVE  ::enes
ORP0028:dag:NOUN:COMMON:SINGULAR:DEFINITE  :NOMINATIVE::en
ORP0028:dag:NOUN:COMMON:SINGULAR:INDEFINITE:GENITIVE  ::s
```

Corresponding DanPO sub-paradigms

`ORP0028.1` **(899 lemmata)**

```
NOUN:COMMON:PLURAL  :INDEFINITE:GENITIVE  ::[0s]
NOUN:COMMON:SINGULAR:DEFINITE  :GENITIVE  ::[!0ns]
NOUN:COMMON:SINGULAR:INDEFINITE:GENITIVE  ::[!s]
NOUN:COMMON:PLURAL  :DEFINITE  :NOMINATIVE::[0n0]
NOUN:COMMON:PLURAL  :INDEFINITE:NOMINATIVE::[0]
NOUN:COMMON:SINGULAR:DEFINITE  :NOMINATIVE::[!0n]
NOUN:COMMON:PLURAL  :DEFINITE  :GENITIVE  ::[0n0s]
NOUN:COMMON:SINGULAR:INDEFINITE:NOMINATIVE::[!]
```

`ORP0028.2` **(331 lemmata)**

```
NOUN:COMMON:PLURAL  :INDEFINITE:GENITIVE  ::[0s]
NOUN:COMMON:SINGULAR:DEFINITE  :GENITIVE  ::[0ns]
NOUN:COMMON:SINGULAR:INDEFINITE:GENITIVE  ::[s]
NOUN:COMMON:PLURAL  :DEFINITE  :NOMINATIVE::[0n0]
NOUN:COMMON:PLURAL  :INDEFINITE:NOMINATIVE::[0]
NOUN:COMMON:SINGULAR:DEFINITE  :NOMINATIVE::[0n]
NOUN:COMMON:PLURAL  :DEFINITE  :GENITIVE  ::[0n0s]
NOUN:COMMON:SINGULAR:INDEFINITE:NOMINATIVE::[]
```

generates the imperative form [tAJ!n] from the stem.

DanPO also enlists lemma-specific compound-formation properties (“glue elements”), such that the dictionary accounts for productive compound morphology.

3 Compliance with the spoken language idiom

Since DanPO is aimed at speech technology, including speech recognition, we needed to ensure the descriptiveness of the dictionary. Therefore we engaged in a cooperation with the DanPASS project (Grønnum, 2005) lead by Dr. Nina Grønnum (Dept. of Linguistics, University of Copenhagen).

The main goal of the DanPASS project (Danish Phonetically Annotated Spontaneous Speech) has been the establishment of Korpus Spontan-Tale (Corpus Spontaneous-Speech) consisting of 57 short monologues (19 speakers performing 3 distinct tasks including a map-guidance task). Each recording was made in the echo free room of Eksperimental-Fonetisk Laboratorum (Experimental Phonetic Lab), and the recordings are of a very high acoustic quality. All recordings were transcribed in a SAMPA-compatible sound alphabet by two phoneticians in parallel. A third phonetician was consulted for each discrepancy found in the two parallel transcription corpora. Spontan-Tale contains about 25,000 tokens annotated with prosodic markup.

Based on the SAMPA-transcription an orthographic-phonetic concordance is derived. In Fig. 3 three selected concordance entries are shown covering some of the types appearing in Fig. 2.

Orthographic-phonetic combinations with few occurrences (C<4) are annotated with transcription references for easy proof reading.

Highly frequent words usually exhibit multiple pronunciation forms and therefore have many alternative entries. An example is the multi-purpose pronoun “der” (there/that) which occurs in 15 phonetic variants, some much more frequent than others. Depending on the grammatical function, “der” is typically pronounced as either (A) or (B), cf. Fig. 3.

(A) is preferred for the expletive use of “der” while (B) is typically used as a locative. When

Figure 3: Sample from the orthographic-phonetic concordance

overgardin		
1	’ÅwágAdi:!n	[m_013_h,t=208]
hedder		
3	heD!á	[m_013_k,t=266] [m_014_k,t=101] [m_033_h,t=11]
2	’heDá	[m_014_k,t=251] [m_016_h,t=187]
5	’heD!á	
46	heDá	
der		
1	deR	[m_019_h,t=145]
49	’deR!	
1	de:!R	[m_021_k,t=22]
4	dV	
35	’dA	
10	’da	
2	dER!	[m_033_h,t=117] [m_033_h,t=137]
1	’dæR	[m_031_g,t=115]
4	deR!	
1	’de:!R	[m_021_k,t=150]
1	d@	[m_007_g,t=41]
1	dæR	[m_031_g,t=48]
223	dA	
1	dER	[m_033_h,t=150]
116	da	

Legend
Each record is indexed by the orthographic form (e.g. “hedder”). Phonetic entries have three fields:

1. no. of occurrences,
2. phonetic representation,
3. transcription references [*filename,time-ref*] (optional)

Figure 4: Pronoun “der”: prototypical phonetic forms

mode	vowel	accent	stød	prototype
(A)	[A]*or* [a]	no	no	[dA]
(B)	[e]	main	yes	[’deR!]

Figure 2: Sample from corpus Spontan-Tale (monologue *m_013_h*)

SAMPA:	sV	’adAed			’Åwá	gA	di:!n		va	’sV	dn,	nå:D	sV	’heD!á	
EnOrt:	så	,er	der	et	,overgardin			+	hvad	s,ådan		noget	så	h,edder	=
Gloss:	then	is	there	an	upper	curtain			what	such		stuff	then	is-called	
Trans:	“then there is an upper curtain or whatever it’s called”														

Legend

SAMPA = transcription using Speech Assessment Methods Phonetic Alphabet
EnOrt = Orthographic rendering enriched with prosodic information (eg. [+] = pause, [=] = hesitation with phonation, [’] = stress)
Gloss = English lexical equivalents
Trans = English translation

Observe that segmentations in SAMPA and EnORT are sometimes in conflict.

such grammatically dependent variation can be detected, multiple phonetic forms are allowed in DanPO.

In general, more frequent phonetic variants are preferred over less frequent, everything else being equal. As explained, in cases where the variation is correlated with the grammatical context, two (or more) alternative phonetic forms are introduced in DanPO and annotated with the according selectional restrictions.

Observe that despite the prototypicality of forms (A) and (B), many other pronunciations of “der” are actually encountered (cf. Fig. 2). Much care must be taken in the selection of prototypical pronunciations for introduction in DanPO. The process of validating preliminary linguistic hypotheses by consulting the transcription files is indeed a labor-intensive one.

Nevertheless, as we argue, there are good reasons to go descriptive. Relying on traditional prescriptive sources (such as dictionaries or linguists’ intuitions) is highly risky. The authors of this paper have often found our personal judgments — even of our own pronunciation — to be misleading. Here we present but a single example. According to one of the major pronunciation dictionaries of Danish (Hansen, 1990), expletive “der” (cf. Fig. 3) is pronounced [dæR], [dA], or [dV] (in that order). Likewise, the locative “der” is [’dæ:!R], [’de:!R] or [’dæR!] (in that order). These pronunciations come close to our own when e.g. presenting “der” to a foreigner.

As the reader may wish to verify (or rather falsify) in Fig. 3, this provides a very poor description of “der” as occuring in actual speech. Only one of the six dictionary forms has any significance in the transcriptions, viz. [dA], while the remaining five forms cover just 6 out of 450 occurrences, or 1.3three dictionary forms accounting for only one single occurrence. As it seems, Danes do not speak by the book.

4 Status and prospects

At the time of writing, the DanPO dictionary contains 87,104 lemmata and morphological information capable of generating 766,474 inflected forms (plus an infinite number of compounds), each of which associated with a phonetic form. Of these, about 1000 are derived using transcription informed phonetics (TIP), as exemplified in section 3 above (lexeme “der”). The remaining phonetic forms are generated using standard phonological rules and methods including traditional hand-coding.

One thousand TIP based phonetic forms may not seem a lot. However, recall that spoken language — especially as occurring in informal situations — recycles the same word types to a much larger extent than is typical for the written genres. Compare e.g. the frequency distribution of two Danish corpora covering spoken language (informal conversations) and written language (newspaper articles), respectively. Each corpus consists of 1,335,000 word tokens (Henrichsen, 2002).

Observe that just 30 word types are needed to cover about half of the transcription corpus while almost 200 types are needed f a similar coverage of the newspaper texts. We have reasons to believe that other languages — maybe all? — show similar distributional patterns (e.g. (Allwood and Henrichsen, 2005), (Leach et al., 2001)).

Figure 5: Word type distribution for spoken and written language

Rank	Spoken lng.cov.		Written lng.cov.	
	Count	Freq.	Count	Freq.
1–10	380,599	28.5%	277,161	20.8%
1–20	549,283	41.1%	412,762	30.9%
1–30	671,223	50.3%	473,882	35.5%
1–100	940,834	70.5%	618,383	46.3%
1–200	1,046,036	78.4%	696,591	52.2%
1–1000	1,197,670	89.7%	876,435	65.6%

A lexicon containing 1000 TIP entries is thus expected to provide TIP based coverage of about 90% of the words occuring in typical ordinary speech.

5 Evaluation, experiments and further development

The heterogeneous status of the dictionary makes it relevant to compare different versions of lemmata and full forms in a systematic way. This would make it possible to judge the quality of the sources in order to choose the direction which development of phonetic dictionaries should take.

As an example, one particularly intriguing lemma (or set of lemmata) is the homograph "der", which has the following phonetic representations in DanPO:

Normative annotation [d2A]

DanPASS transcriptions

Freq	SAMPA	DanPO
45	`'deR?`	[d2eR!]
33	`'dA`	`[d2A]`
234	`dA`	`[dA]`
100	`da`	`[da]`

"Editor's choice" [d2A]

The lemma occurs very differently whether pronounced in stressed or unstressed versions. The manual editor suggests the normative choice. We plan on conducting systematic naturalness judgements of phrases containing lemmata with alternative phonetics.

Furthermore, we suggest that segments, frequency (F0) contours, and segmental durations be refined by use of the Segment Editor developed by Peter Rossen Skadhauge. The Segment Editor, whose main functionality is depicted in Fig. 6, facilitates editing of the segmental quality, duration and frequency for every segment in an utterance. Segments may be inserted, changed, or deleted at random places in the utterance. Since frequency (F0) is shown as horizontal sliders, the graphical picture of all the frequency sliders may be seen as an intonation curve for the utterance, which may be altered by adjusting the sliders individually.

The example shows the state of the editor just having loaded a raw phonetic sequence corresponding to the text "Der er ikke noget at gøre ved det".

The Segment Editor may be used to facilitate improvement of utterance synthesis in the following ways:

- Experts' hand-tuning of parameters
- Informants' hand-tuning of parameters by negotiation.

We are going to set up experiments where informants negotiate parameter values for determination of optimal rendering of synthesis. These parameters may, in turn, be used as a basis of machine-learning intonation patterns for spoken language.

6 Concluding remarks

The first version of DanPO is finished. Judgments of DanPO's potential for speech technological improvements are preliminary, but have shown the DanPO lexicon to significantly improve the naturalness of the Danish Synthetic Voice (Henrichsen, 2004). In a pilot experiments, we shall present a panel of native speakers of Danish with samples of synthetic speech in two variants, with and without TIP based versions of DanPO, keeping everything else unchanged, such as lexical content, fundamental frequency contour, timing, and voice quality.

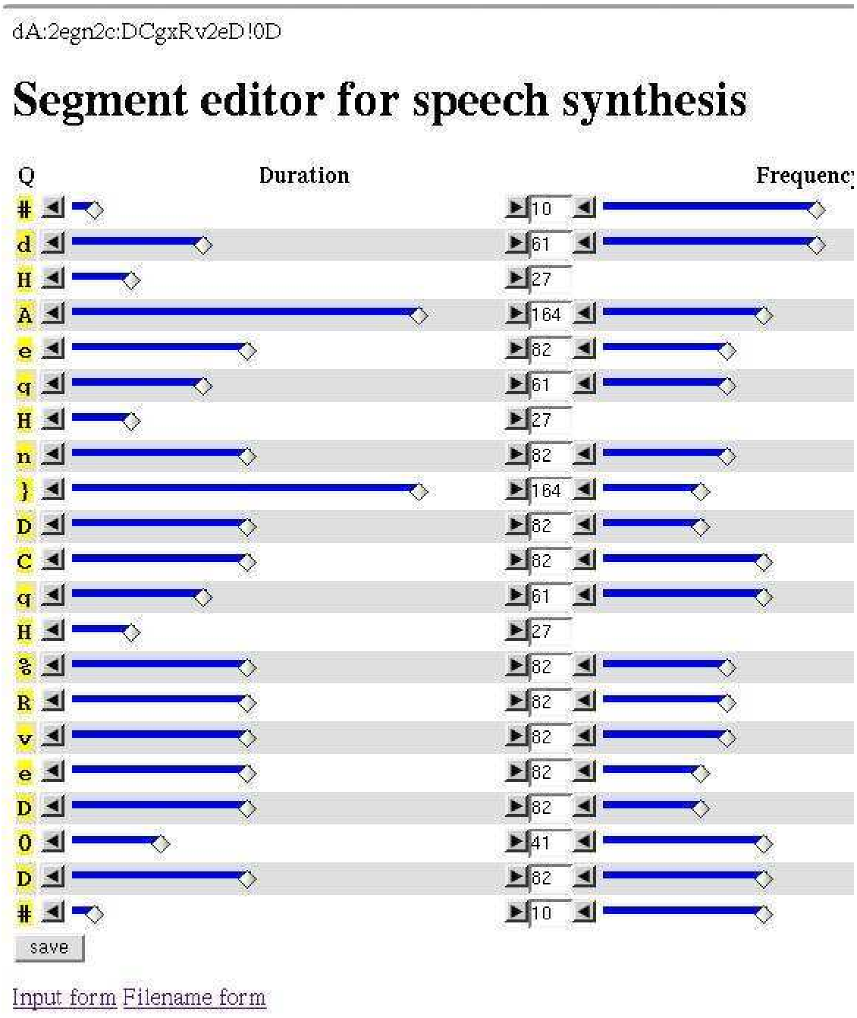
References

Jens Allwood and Peter Juel Henrichsen. 2005. Swedish and Danish, spoken and written language - a statistical comparison. *International Journal of Corpus Linguistics*, 10(3):367–399.

Anna Braasch. 2003. Sto: A lexical database of Danish for language technology applications. *elsnews*, 12(3), Autumn.

Nina Grønnum. 2005. Danish phonetically annotated spontaneous speech. `http://www.cphling.dk/~ng/danpass.html`.

Figure 6: The Segment Editor



dA:2egn2c:DCgxRv2eD!0D

Segment editor for speech synthesis

Q	Duration		Frequenc
#		10	
d		61	
H		27	
A		164	
e		82	
q		61	
H		27	
n		82	
}		164	
D		82	
C		82	
q		61	
H		27	
%		82	
R		82	
v		82	
e		82	
D		82	
0		41	
D		82	
#		10	

Input form Filename form

Peter Molbæk Hansen. 1990. *Dansk Udtale*. Gyldendal, Nordisk Forlag.

Peter Juel Henrichsen. 2002. Some frequency based differences between spoken and written Danish. In *Gothenburg Papers in Theoretical Linguistics*, volume 88.

Peter Juel Henrichsen. 2004. The twisted tongue. tools for teaching Danish pronunciation using a synthetic voice. In Peter Juel Henrichsen, editor, *Copenhagen Studies in Language*, volume 30, pages 95–111. Copenhagen Business School Press.

Geoffrey Leach, Paul Rayson, and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English*. Longman.

Peter Rossen Skadhauge and Peter Juel Henrichsen. 2005. Danish phonetic-orthographic lexicon. `http://www.id.cbs.dk/~prs/danpo`.

Larry Wall, Tom Christiansen, and Jon Orwant. 2000. *Programming Perl*. O'Reilly, 3rd edition, July.

J. C. Wells. 1997. Sampa computer readable phonetic alphabet. In D. Gibbon, R. Moore, and R. Winski, editors, *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin and New York. Part IV, section B.