The DK-CLARIN corpus of speech and writing

Peter Juel Henrichsen
Center for Computational Modelling of Language (CMOL)
Copenhagen Business School
pjh.isv@cbs.dk

Abstract

Ever since its creation in the mid-nineties, the Danish PAROLE corpus has played a key role as a reference corpus in the Danish corpus-linguistic research. Over the last decade, PAROLE has been continuously developed at CMOL and elsewhere by adding several new annotation dimensions to the original PoS annotated orthographic text. Today approximately 10 annotation tiers exist for the 250k token corpus of mixed text genres. This document presents PAROLE's speech-related annotation tiers: a sound track covering 100k tokens read aloud by one male speaker along with phonetic transcription, prosodic mark-up, and acoustic tracings for F0, intensity, and voicedness (Harmonicity-to-Noise ratio). The annotated PAROLE will be made available to researchers through the DK-CLARIN web-services (cf. http://dkclarin.ku.dk/).

1. The Danish PAROLE corpus

The Danish PAROLE corpus was established in the mid-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 creating a (by that time's standards) large text corpus. The corpus was to be balanced with respect to text types, that is, compiled from various text sources including newspaper, magazine, and journal. A substantial subpart of this corpus (at least 250k words) was to be manually annotated for part-of-speech. The annotation work was carried out mainly by Thomas Bilgram and Britt Keson, supervised by Ole Norling Christensen.

Figure 1 below shows a sample from the first section of the PAROLE text body in the original SGML markup, representing the newspaper quote “De mener, at Folkekongressen skal give præsidenten diktaroriske beføelelser.” (*They believe that the Duma should give the President dictatorial powers*).

```
<W lemma="de" msd="PP3[CN]PN-NU">De</W>
<W lemma="mene" msd="VADR=----A-">mener</W>
<W lemma="," msd="XP">,</W>
<W lemma="at" msd="CS">at</W>
<W lemma="folkekongres" msd="NCCSU==D">Folkekongressen</W>
<W lemma="skulle" msd="VADR=----A-">skal</W>
<W lemma="give" msd="VAF-=----A-">give</W>
<W lemma="præsident" msd="NCCSU==D">præsidenten</W>
<W lemma="diktaroriske" msd="XX">diktaroriske</W>
<W lemma="beføelelser" msd="XX">beføelelser</W>
<W lemma="." msd="XP">.</W>
```

Figure 1. PAROLE sample

Observe that the PAROLE corpus represents the original texts as-were, thus including intentionally a rather substantial amount of typos and other errors. Examples are the words in **bold** typeface which should probably have been "diktatoriske beføjelser" (represented as such in the English translation above). Their status as typos are designated by the tag **XX** while the other tokens have of course more informative PoS.

1.1 PAROLE's morpho-syntactic annotation

PAROLE's PoS annotation uses a cleverly designed annotation format allowing for easy transfer of morpho-syntactic information across language boundaries. By way of example, "2nd person pronouns, plural forms" can be searched for across all European PAROLE corpora using the same search template. PAROLE PoS tags consist of a string of characters, each specifying a morphological feature/value. The features are ordered from more-general to more-specific. Fig. 2 presents an example, the PAROLE-tag for "husenes" (*the-houses'*) **NCNPG==D**.

```
NCNPG==D       "husenes"
-----------------------------------------------------
N          "noun"     hus-
 C         "common"
  N        "neuter gender"
   P       "plural"        -e-
    G      "genitive"                    -s
     =     (void for Danish)
      =    (void for Danish)
       D   "definite"                -ne-
-----------------------------------------------------
```

Figure 2. Hierarchical style in PAROLE tags

The full PoS definition table is included in the appendix. More information on the PAROLE tag set can be found in Keson (1999) and at

http://www.elda.org/catalogue/en/text/doc/parole.html (pan-European)
http://korpus.dsl.dk/paroledoc_dk.pdf (Danish)

The original PAROLE source containing 250,000 word-like tokens (290,600 tokens in total including punctuation and other non-alphanumeric text elements) is now publicly available for non-commercial uses. The SGML-file can be downloaded free of charge from the web-pages of Det Danske Sprog og Litteraturselskab (www.dsl.dk), together with manuals and descriptions in Danish and English language.

1.2 Status (June 2009)

Since the conclusion of the European PAROLE project, the annotation work has been continued at CMOL. Today PAROLE is a poly-dimensional corpus structure including these annotation dimensions (in various states of completion):

- Syntactic structure (dependency based, aka. Copenhagen Dependency Treebank)
- Rhetorical structure (RST based)
- English translation (manually prepared)
- Russian translation (do.)
- Tamil translation (do.)
- Phonetic annotation (automatically generated, manually revised)
- Prosodic annotation pitch marking (manually prepared)
- Sound track (reading by one male speaker, sound studio quality, two parallel channels: chin mounted and fixed mics, 44.1kHz sample rate)
- Acoustic measurement (each 5 ms window) for
  - *F0 (fundamental frequency)*
  - *intensity*
  - *voicedness*

One of the most recent developments is the inclusion of a sound track along with phonetic transcription as well as acoustic measurements for various parameters, most notably F0 (fundamental frequency), intensity, and voicedness (HNR, harmonicity-to-noise-ratio, cf. http://www.praat.org).

2. The recordings

One male Danish speaker (the author) read the first 100,000 tokens of the PAROLE text in low-echoic surroundings (the speech studio Hedehuset, Nykøbing S) using professional quality sound equipment. The set-up included two separate channels: a chin mounted microphone and a stationary (cardioid) capacitor microphone. The sampled signal (2×44.1kHz) was streamed directly to a laptop harddisk using a USB-port.

For reasons of economy, the recording sessions had to be performed by a single person filling both roles of technician and speaker. Each reading session (covering 5k tokens on average) was therefore performed non-stop without attempts to repair erroneous readings. Reading errors typically occurred at irregular text instances (typos, broken syntax, non-standard abbreviations, foreign proper names, esoteric symbols and numerals etc.), but of course simple slips-of-the-tongue also occurred. In such cases, a correction note was made in the reading manuscript for later recovery. In total, 5% of all readings were judged by the reader as non-satisfactory during the recording sessions. To this came errors detected in later stages, including some instances of acoustic oversteering, resulting in a total of 15% retakes (followed by a small number of second-generation retakes as well).

3. Post-recording operations

3.1 Beep segmentation

At recording time, each text section read was followed by a beep inserted by the speaker (using a kitchen stopwatch). Based on these beeps, the sound file were later segmented automatically. Due to the good acoustic quality (low background noise and carefully monitored recording level, resulting in a high signal-noise ratio), this could be done with very high confidence, better than 1 error in 2000 beeps. A simple tool was implemented for verifying the alignment of the original text and the spoken reproduction.

3.2 Acoustic analysis

For each 5 ms window, various acoustically related data points were derived using algorithms included in the PRAAT open source platform for phonetic analysis (http://www.praat.org) – most notably for pitch (F0), voicedness (HNR, harmonicity-to-noise), and intensity. Details are found in Henrichsen et al 2009 and elsewhere. The acoustic data are available as tiers conforming to standard PRAAT textgrid format.

3.3 Phonetic transcription

The PAROLE reading corpus was transcribed by two advanced MA-level phonetics students (KUA), supplemented by a professional phonetician (AU). A slightly modified variant of standard SAMPA (http://www.phon.ucl.ac.uk/home/sampa/danish.htm) was used. As seen, most symbols are identical to their SAMPA equivalents. Digits and other non-alphabetic characters are however avoided since these tend to co-incide with symbols defined in e.g. search languages and formalisms such as EBNF and Regular Expressions.

Full vowels:	`i e E z a A y q Q x u o c C X`
Schwa:	`0 -`
Consonants:	`b d D f g h j J k l m n N p r R s S t v w`
Diacritics:	`2 : ?`

Deviations from standard SAMPA:

Symbol		*SAMPA equivalent*
`0`	schwa	`@`
`-`	assimilated schwa	`@`
`z`	as in "las"	`{`
`q`	as in "løs"	`2`
`Q`	as in "løn"	`9`
`c`	as in "lund"	`O`
`x`	as in "grøn"	[*missing*]
`X`	as in "vor"	`Q`
`C`	as in "som"	`Q` (?)
`S`	as in "sjov"	`sj`
`2`	main stress	`"`
`!`	stød	`?`
`-`	stød	`?`

As mentioned, the texts of PAROLE are not highly polished literary products, but rather quickly produced news articles, reader's comments, unpretentious entertainment, technical specifications for cars, sports results, financial notes, and so on. Typos and strange names are frequent, and many text sections do not at all resemble syntactically well-formed sentences.

"MÅL OG VÆGT : Længde/<bredde/<højde : 432/166/142 cm."
"Size and weight: length/width/height: 432/166/142 cm."

"(..) Ständige Musterausstellung , Poststrasse 1 , W-7530 Pforzheim ."
(German postal address)

"I går blev Kosan Teknova solgt fra , og tilbage i Kosan Holding er kun Crisplant ."
Yesterday, K.T. was sold, and remaining in K.H. is only Crisplant.

"Fakta om Seat Toledo 2,0 GLX OPBYGNING : 4/5-personers sedan"
"Facts about the Seat Toledo 2.0 GLX CONSTRUCTION : 4/5-person sedan" ,

Since we decided to represent the PAROLE corpus in full, none of these cases are omitted – at the price of some tongue twisting readings at times. For non-Danish words, we tried to keep the readings within the Danish phonetic realm, as far as possible. In the transcription of some English and other non-Danish words, however, we had to allow `T`, `W`, and `L` ("th at", "w ood", "gi rl", respectively).

4. Concluding remarks

We believe that the cumulative strategy in corpus development is a highly rewarding one. Though PAROLE is, by current standards, a rather small corpus, small can be beautiful even in corpus linguistics. For most purposes, the value of a text corpus is not doubled by adding 100% more text. However, in contrast, providing the corpus with a separate new annotation tier for the same text – adding again a similar amount of new data – may allow totally new uses of the corpus. Using various combinations of the PAROLE tiers we have conducted several experiments which would not otherwise have been possible, including corpus-based learning of stress assignment (Henrichsen 2001), data-driven phonetic lexicography (Henrichsen et al 2005), and recycling of phonetic resources for cognate languages (Henrichsen 2007a, 2007b).

We hope that the new PAROLE resource will contribute to new developments, in particular analyses of spoken Danish including prosody and intonation. We feel confident that PAROLE, with its tight alignment of text, phonetics, and acoustic measurements will become an even more useful material for phonetic and linguistic studies in general, as well as an essential data source for speech technological development.

References

Henrichsen, Peter Juel (2007a) *A Norwegian letter-to-sound engine with Danish as a catalyst;* NODALIDA-07 (Tartu)

Henrichsen, Peter Juel (2008) *NoTa – nu med lydskrift;* in J.Bondi Johannessen & al (eds) (2008) *"Språk i Oslo. Ny forskning omkring talespråk"*, Novus forlag, Oslo

Henrichsen, Peter Juel; Peter Rossen Skadhauge (2005) *DanPO – a Danish phonetic-orthographic dictionary for speech technological integration;* ICOIL-2005 (Chennai)

Henrichsen, Peter Juel; Thomas Ulrich Christiansen (2009) *Fishing for meaningful units in connected speech; ISAAR-2009*

Keson, B. (1999) *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus*; Det Danske Sprog- og Litteraturselskab (se http://korpus.dsl.dk/e-resurser/parole-korpus.php)

Henrichsen, P. J. (2001) *Transformation Based Learning of Danish Stress Assignment*; Eurospeech-2001, Aalborg, Danmark

Grønnum, Nina (1998) *Fonetik og Fonologi - Almen og Dansk*; Copenhagen: 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

Appendix. Definition table for the Danish PAROLE taxonomy

Legend

- **.** *value underspecified*
- **=** *feature not defined for Danish (but for some other European language)*
- **-** *feature not instantiated for this particular form*
- **:** *tag truncated (avoiding repetition of information present elsewhere)*

We allow some redundancy in the table to enchance legibility.

ANPCSU=IU	adjective normal positive common singular unmarked-case indefinite unmarkeduse
ANP.SG=DU	adjective normal positive singular genitive definite unmarked-use
ANP.SU=DU	adjective normal positive singular unmarked-case definite unmarked-use
ANP.SU=IU	adjective normal positive singular unmarked-case indefinite unmarked-use
ANP.P:	adjective normal positive plural (...)
ANPN:	adjective normal positive neuter (...)
ANP---=-R	adjective normal positive adverbialuse
ANC:	adjective normal comparative (...)
ANS:	adjective normal superlative (...)

AC---G=--	adjective cardinal genitive
AC---U=--	adjective cardinal unmarkedcase
AO---G=--	adjective ordinal genitive
AO---U=--	adjective ordinal unmarkedcase
CC	conjunction coordinative
CS	conjunction subordinative
I	interjection
NCCSG==D	noun commonnoun common-gender singular genitive definite
NCCSG==I	noun commonnoun common-gender singular genitive indefinite
NCCSU==D	noun commonnoun common-gender singular unmarked-case definite
NCCSU==I	noun commonnoun common-gender singular unmarked-case indefinite
NCN:	noun commonnoun neutergender (...)
NCCP:	noun commonnoun commongender plural (...)
NP--G==-	noun proper genitive
NP--U==-	noun proper unmarkedcase
PC--PG---	pronoun coordinative plural genitive
PC--PU---	pronoun coordinative plural unmarked-case
PD-CSU--U	pronoun demonstrative common singular unmarkedcase unmarkedstyle
PD-CSU--O	pronoun demonstrative common singular unmarkedcase obselete
PD-CSG:	pronoun demonstrative common singular genitive (...)
PD-N:	pronoun demonstrative neuter (...)
PD-.P:	pronoun demonstrative plural (...)
PI:	pronoun indefinite (...)
PO1CSUSNU	pronoun possessive firstperson common singular unmarked-case singular non-reflexive unmarkedstyle
PO1CSUPNF	pronoun possessive firstperson common singular unmarked-case plural non-reflexive formal

PO1:	pronoun possessive firstperson (...)
PO2:	pronoun possessive secondperson (...)
PO3CSUSYU	pronoun possessive thirdperson common singular unmarkedcase singular reflexive unmarl
PO3:	pronoun possessive thirdperson (...)
PP1CSN-NU	pronoun personal firstperson common singular nominative nonreflexive unmarkedstyle
PP1CSU-.U	pronoun personal firstperson common singular unmarkedcase unmarkedstyle
PP1CP:	pronoun personal firstperson common plural (...)
PP2:	pronoun personal secondperson (...)
PP3CSN-NU	pronoun personal thirdperson common singular nominative nonreflexive unmarkedstyle
PP3NSU-NU	pronoun personal thirdperson neuter singular unmarkedcase nonreflexive unmarkedstyle
PP3..U-YU	pronoun personal thirdperson unmarkedcase reflexive unmarkedstyle
PP3.P:	pronoun personal thirdperson plural (...)
PT:	pronoun interrogative-relative (...)
RGA	adverb general absolutesuperlative
RGC	adverb general comparative
RGP	adverb general positive
RGS	adverb general superlative
RGU	adverb general unmarkedcomparison
SP	adposition preposition
U=	unique (*“at”, “der”*)
VADR=----A-	verb main indicative present active
VADR=----P-	verb main indicative present passive
VADA:	verb main indicative past (...)
VAF-:	verb main infinitive (...)
VAG-=SCI--U	verb main gerund singular common indefinite unmarkedcase
VAM-=------	verb main imperative
VAPR=---R--	verb main participle present adverbialuse

VAPR=...A-U	verb main participle present unmarkedcase
VAPR:	verb main participle present (...)
VAPA=SCDA-U	verb main participle past singular common definite unmarkedcase
VAPA:	verb main participle past (...)
VEDR=----A-	verb medial indicative present active (*“synes”, “slås”*)
VE:	verb medial (...)
XA	residual abbreviation
XF	residual foreign-word
X:	residual (...)