A Dictionary of Spoken Danish

Carsten Hansen & Martin H. Hansen

Keywords: *lexicography, speech corpus, pragmatics, conversation analysis.*

Abstract

The purpose of this project is to establish a dictionary of spoken Danish, titled Ordbog over Dansk Talesprog (ODT). Through the use of extensive empirical data, it is the aim of the project to convey the latest knowledge of spoken language to the broad public. ODT combines existing and new research based primarily on qualitative methods with the quantitative analysis of a corpus of spoken language. The result of this combined method will be made available to the public through the development of a web-based dictionary of spoken Danish.

ODT is a project of the Centre for Language Change in Real Time (LANCHART) at the University of Copenhagen. Building on a large corpus of spoken language consisting primarily of sociolinguistic interviews, recorded from 1978 – 2010 and consisting of almost 7 million transcribed tokens, we are working on a dictionary portal. We inscribe the project into a tradition of significant national dictionaries, namely the Dictionary of the Danish Language (1918 – 1956) and The Danish Dictionary (2003 – 2005). Both were published by the Society for Danish Language and Literature, which is one of our foremost institutional cooperating partners along with the Danish Language Council.

The ODT project pursues two spheres of action. One lets the editors conduct research of their own, both in the field of spoken-language research in line with the other activities at the LANCHART Centre, and in the new field of spoken-language lexicography. In this way the editors, future dissertation writers, and Ph.D. students working on the project will produce new knowledge. The other sphere of action concerns conveying this knowledge to the public. We see it as our job not only to promote and expose the research activities of the editors themselves and the other LANCHART researchers, but also to pass on knowledge and research on spoken language gained outside of the Centre.

The user segment of ODT consists of two groups. The primary recipient is the linguistically curious layperson interested in spoken language; the secondary recipient is the research oriented user. Both groups will benefit from a web portal which allows fast access, is segmentally differentiated (i.e., relevant), has a high level of service, is free of advertising, and is free to use.

ODT is designed as a web-based dictionary portal with a possibility for parallel comparable searches in a corpus of written Danish (KorpusDK) and in a dictionary mainly based on written Danish (The Danish Dictionary).

Theoretical work on ODT consists in elaborating on well-established lexicographic methods and exploring the possibilities for transferring them into a dictionary of spoken language. The practical work consists of actual dictionary compilation: searching, editing, storing, and presenting the corpus data.

1. Introduction

The purpose of this project is to establish a dictionary of spoken Danish, titled *Ordbog over Dansk Talesprog* (ODT). Through the use of extensive empirical data, it is the aim of the project to convey the latest knowledge of spoken language to the broad public. ODT combines existing and new research based primarily on qualitative methods with the quantitative analysis of a corpus of spoken language. The result of this combined method will be made available to the public through the development of a web-based dictionary of spoken Danish.

ODT is a project of the Centre for Language Change in Real Time (LANCHART) at the University of Copenhagen. Building on a large corpus of spoken language consisting primarily of sociolinguistic interviews, recorded from 1978 – 2010 and consisting of almost 7 million transcribed tokens, we are working on a dictionary portal. We inscribe the project into a tradition of significant national dictionaries, namely the *Dictionary of the Danish Language* (1918 – 1956) and *The Danish Dictionary* (2003 – 2005). Both were published by the Society

for Danish Language and Literature, which is one of our foremost institutional cooperating partners along with the Danish Language Council.

The ODT project pursues two spheres of action. One lets the editors conduct research of their own, both in the field of spoken-language research in line with the other activities at the LANCHART Centre, and in the new field of spoken-language lexicography. In this way the editors, future dissertation writers, and Ph.D. students working on the project will produce new knowledge. The other sphere of action concerns conveying this knowledge to the public. We see it as our job not only to promote and expose the research activities of the editors themselves and the other LANCHART researchers, but also to pass on knowledge and research on spoken language gained outside of the Centre.

The user segment of ODT consists of two groups. The primary recipient is the linguistically curious layperson interested in spoken language; the secondary recipient is the research oriented user. Both groups will benefit from a web portal which allows fast access, is segmentally differentiated (i.e., relevant), has a high level of service, is free of advertising, and is free to use.

ODT is designed as a web-based dictionary portal with a possibility for parallel comparable searches in a corpus of written Danish (*KorpusDK*) and in a dictionary mainly based on written Danish (*The Danish Dictionary*).

Theoretical work on ODT consists in elaborating on well-established lexicographic methods and exploring the possibilities for transferring them into a dictionary of spoken language. The practical work consists of actual dictionary compilation: searching, editing, storing, and presenting the corpus data.

2. Fact boxes

The user interface has two levels. Apart from regular dictionary entries with audible sound clips, a number of 'fact boxes' will be written and cross-referenced to the relevant entries. In these boxes, the editors and various guest authors will concisely characterize selected linguistic phenomena according to a 'box manual', as follows:

The presentation *must*

- characterize a spoken language phenomenon
- be founded on corpus examples
- contrast speech with writing
- be addressed to both of the user segments.

The presentation *can*

- be based on data other than the LANCHART Corpus (this must then be explicitly indicated)
- include previous research in the phenomenon.

The subject of a fact box can be almost any aspect of speech. It can be a lexeme, a function group or any other relevant subject, such as laughter, pauses, and new constructions like the suffix *–agtig* ('-like'). The editorial staff is constantly looking for subjects as well as guest authors for boxes.

3. Pilot project on interjections

The ODT project is currently in an initial explorative state, both theoretically and methodologically. Taking advantage of our speech corpus consisting mainly of sociolinguistic interviews, we are carrying out a pilot project on interactional tokens, which are annotated as interjections in our corpus. Our corpus has been automatically part-of-speech (PoS) annotated; in order to avoid possible mistakes in the PoS annotation, we supplement the list of interjections generated from the corpus with an interjection inventory from *The Danish Dictionary*, which is based on a corpus of 40 million tokens from mainly written texts published around the year 2000. In this explorative state, our praxis is somewhat lax in regard to distinguishing between the PoS categories 'interjection', 'onomatopoeia', and 'particle'. Schwitalla (2003) and Fiehler (2005) suggest a functional co-category for these three categories named 'Gesprächspartikeln' ('speech articles').

We perform two different but complementary procedures which supplement each other. In the first procedure, we go through the two previously mentioned interjection inventories in a semasiological way and allocate every single candidate to a 'function group', asking what kind of job the candidate in question fulfills in the conversation. From Adolphs (2008), we have picked up the term 'functional profile', which indicates the total sum of different but supplementary functions of the lexeme. Schwitalla (2003: 157) and Fiehler et al. (2004: 204ff) have inspired us to develop an enhanced typology of the function group, while knowing full well that Schwitalla's and Fiehler's lists are orientated towards sequential categorization (turn-taking in a Conversation Analysis (CA) sense). In the current state, our categorization is also oriented toward content and speech act.

In the second procedure, we supplement the developed function groups with candidates gleaned from the semasiological analytic procedure. We check whether the functional profiles can add new functions to the total inventory of function groups. Tentative examples of the function groups are 'regret', 'confirmation', 'worry', 'acknowledgement', 'greeting', 'surprise (positive)', 'surprise (negative)', 'self correction', 'hesitation', 'skepticism', and 'disgust'.

4. The dictionary entries

Each dictionary entry will contain a cross reference to the function group to which the lemma belongs. The full list of functions that the lemma can perform – its functional profile – will be supplemented with corpus-based information about the distribution of its functions in different kinds of discourse.

Example 1 presents an early (translated) version of an entry on the interjection *av* ('ouch'), illustrating the way corpus instances are assigned to different function groups. The examples are made anonymous in accordance with the LANCHART policy of guaranteeing full anonymity to informants.

Example 1
Lemma: *av* ('ouch')

av[1] (surprise)
Instances in the corpus: 40
Usage:
av: used to express surprise or other (often unpleasant) feeling: *Av, det dur ikke* ('Ouch, that doesn't work') (Vinderup 2006)

av for satan/søren/pokker/dælen, av min arm: used as a mild profanity to express annoyance or surprise: *Av for satan, en historie altså* ('Ouch, hell, what a story') (BySoc 2007)
Function group: surprise (negative), regret

av^{2} (minimal response)
Instances in the corpus: 27
Usage:
av: used to acknowledge what another speaker says: *Ja… ja… av… … laver I andet* ('Yes… yes… ouch… …do you do other things') (Odder 1988)
Function group: minimal response

av^{3} (acknowledgement)
Instances in the corpus: 1
Usage:
av ja: used to agree, acknowledge, admit: *Det er jo det… av ja, det er varmt* ('That's right… ouch yes it's hot') (Næstved 2006)
Function group: acknowledgement/agreement

The three functions of *av* in the example illustrate how different kinds of functions can be performed by the same lemma: **av**^{1} and **av**^{3} express an emotion, while **av**^{2} is used to organize speech. This example demonstrates how an established lexicographic procedure can be combined with the concept of function groups.

5. Another example: Laughter

In our inventory of interjections we find, among others, the text string *ha*. In the LANCHART corpus transcription, all sounds that sound like laughter have been given as *ha*. On the plus side this means that we can quickly access all 60,000 instances of laughter in the corpus; the problem remains, however, to identify and distinguish between different kinds of laughter. In other words, the editorial task at hand is to develop a function based laughter typology and assign the corpus instance to its types.

It is evident that laughter is more than just an automatic response to humor. CA research on laughter has shown that laughter is sequentially organized, that laughing can (and most often does) invite shared laughter, and that different kinds of response to laughter can establish and define social roles (Journal of Pragmatics 42 (2010)).

It has also been shown (Glenn (2003: 48)) that some instances of laughter can only be explained if the laughter is seen as referring to a 'laughable'. In other words, laughter does not only function as a way to organize conversation; it is also used to characterize the 'laughable' as peculiar in some way – funny, strange, or unexpected. The 'laughable' may be the subject of the conversation, it may be a participant (or non-participant) in the conversation, or it may be (some part of) the situation in which the conversation takes place. By referring to a 'laughable', then, laughter can be said to function as a constative or similar speech act.

Thus, laughter seems to fit into the same functional frame as interjections like *av* in the previous example. It has several functions, some of which have to do with constructing and maintaining the conversation, and some of which have to do with expressing or modifying the speaker's (or laugher's) intended meaning.

6. A functional profile of laughter

The qualitative approach of laughter research within CA has revealed a number of possible laughter functions. Combined with a quantitative corpus approach, the dictionary entry on laughter will enable us not only to assess the frequency of these functions, but also to describe their discursive distribution. For the purpose of the provisional presentation in this section, no distinction is yet made between the possibly different physiological kinds of laughter (giggling, single laugh particles, 'smiling voice', and smiling have all been identified as phenomena related to laughter).

The LANCHART corpus annotation allows for a description on two different discursive levels: a sociolinguistic description of the use of any linguistic unit (including laughter) with regard to age, gender, geography, social class, and distribution in real time over three decades, as well as a discourse context annotation that breaks the conversation into various genres, interaction types, speech act types, and activity types.

To illustrate the potential value of adding discursive distribution to the dictionary entry, we make some observations about the distribution of laughter. Note that these observations will be more informative when we can distinguish between the different laughter functions.

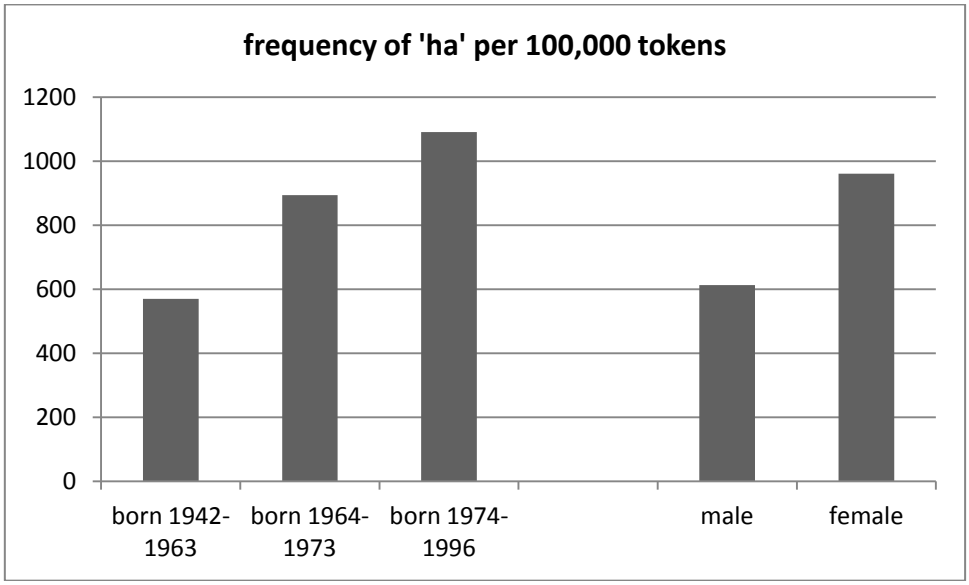

Figure 1. Laughter frequency distributed over age and gender.

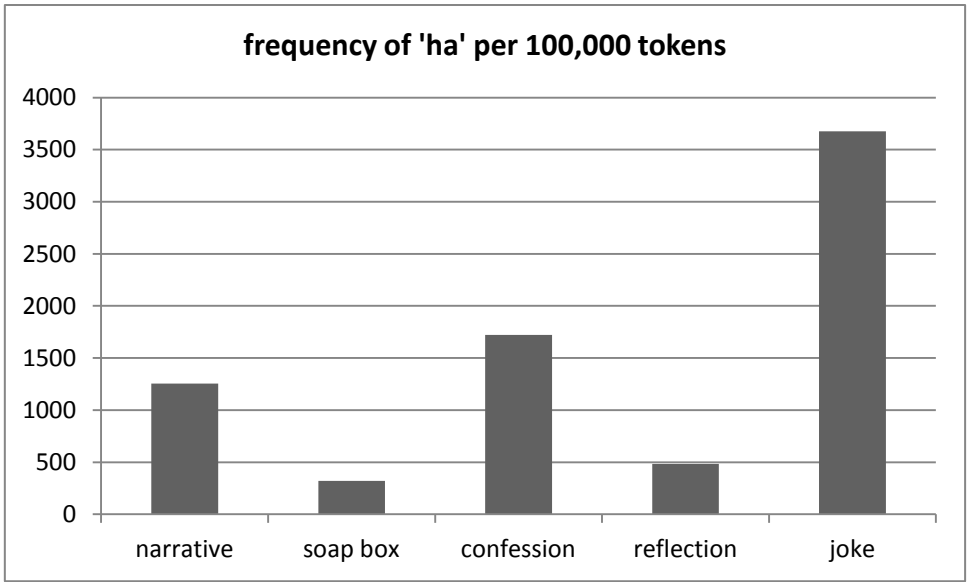

Figure 2. Laughter frequency distributed over selected genres.

Figure 1 seems to show that younger people laugh more often than older people, and that women laugh more often than men. Figure 2 shows that the frequency of laughter varies a great deal between different genres. A detailed description of the genre annotation is available on the LANCHART website (Kodningsmanual til diskurskontekstanalyse: 36-53).

While these graphs are only a shallow reproduction of the frequencies found in our speech corpus that call for further analysis and interpretation, they do suggest that it is worthwhile to implement discursive distribution in a dictionary.

7. Concluding remarks

Although the work presented in this article is preliminary in nature, the underlying concept of developing a dictionary on the basis of the LANCHART corpus is fertile. These naturalistic speech data represent a stratified sample of selected sociolinguistic and discursive parameters.

References

A. Dictionaries

Den Danske Ordbog. http://ordnet.dk/ddo.
KorpusDK. http://ordnet.dk/korpusdk/.
Ordbog over det danske Sprog. http://ordnet.dk/ods/.

B. Other literature

Adolphs, S. 2008. *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse.* Amsterdam: John Benjamins.
Fiehler, R. 2005. 'Die Gesprächspartikel'. In *Duden Band 4*. (Seventh edition). Mannheim: Duden Verlag, 601–606.
Fiehler, R. (et al.) 2004. *Eigenschaften gesprochener Sprache*. Tübingen: Gunter Narr Verlag.
Glenn, P. 2003. *Laughter in Interaction*. Cambridge: Cambridge University Press.

Kodningsmanual til diskurskontekstanalyse (aka. iiv-analyse). **2011.** http://dgcss.hum.ku.dk/aarsberetninger/rapporter/IIV-kodningsmanual__opdateret_januar_2011___2_.pdf/.

Schwitalla, J. 2003. *Gesprochenes Deutsch. Eine Einführung.* Berlin: Erich Schmidt Verlag.

Vöge, M. and J. Wagner 2010. 'Social Achievements and Sequential Organization of Laughter.' *Journal of Pragmatics* 42: 1469–1473.