Frans Gregersen
THE LANCHART Corpus of Spoken Danish, Report from a corpus in progress

Background

The LANCHART corpus of spoken Danish in real time was collected for various purposes, primarily the purpose of evaluating how investigations of variation in contemporary language may be used to assess rapid and progressing language change. Since the pioneering work of William Labov in the mid 1960ies (Labov 1966 cf the new edition 2006), the standard method has been to collect spoken language data from age and gender (and most often also social class) stratified samples of informants. Such informant samples may document language *change in apparent time.* The hallmark is the typical pattern of younger persons having other variable values than older persons. The apparent time hypothesis as an indicator of language change presupposes that individuals after a certain period remain stable in their use of the variables in question. This particular presupposition has been under attack from two different angles:

The first one concerns so-called *age grading*: *If* there are certain periods in any person's life which is typically marked by a particular use of variables and *if* this use is typically given up once the person progresses to another stage, *then* the variables in apparent time may be misleading as to whether a change is actually in progress or not. The use of the young people will invariably be different from that of the older persons but once the young people become older they will leave the 'young use' of the variable behind while other young persons will take over.

The other problem is more fundamental. Gillian Sankoff and her colleagues have in recent years been investigating what they call *language change across the life span* (Sankoff 2005). If no person remains stable across the life span, then again *apparent time* may be a misleading research strategy. And if only certain variables change across the life span, it will be important to know which ones and what may be the significant characteristics of such variables.

Two significant problems to be solved

Before we may even come close to solving these problems which all of them have to do with the relationship between synchronic variation and diachronic change, we have to solve at least two other methodological problems, viz. *the style problem* and *the comparability problem*. The two problems are closely related.

The style problem

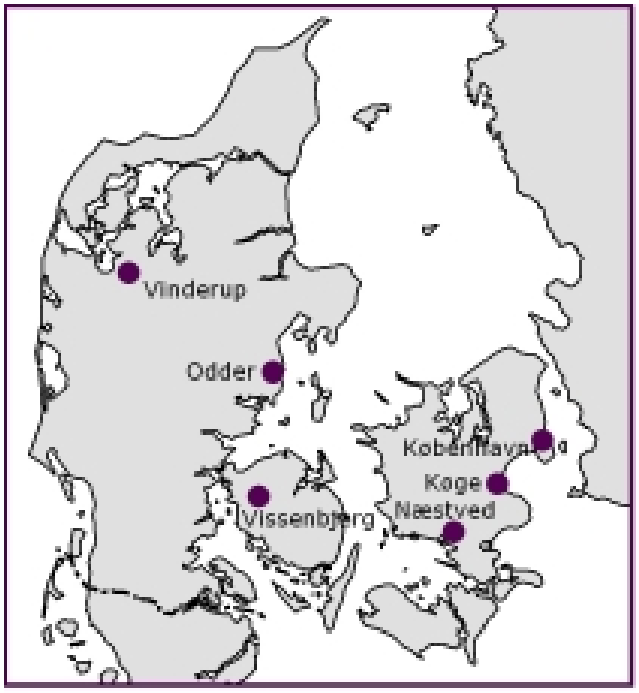
We know that individuals are not stable in their use of variants across situations (Milroy and Gordon 2003: 200ff, Meyerhoff 2006: 28-51). The standard method of collecting data for sociolinguistic research almost invariably involves the use of sociolinguistic interviews. In such interviews the investigator tries to lead the informant towards the use of what has variably been called 'the vernacular', 'casual style' or a host of other names denoting the style supposedly used when the interviewer is not present (cf the Observer's paradox, Meyerhoff 2006: 38f). There are various prescriptions for sociolinguistic interviews (Labov 1984 being the most thorough) but they all involve a change within *the interview frame* between more formal sequences and more relaxed ones. The crucial problem is how to delimit these sequences in a principled way so that they may be contrasted.

This is central for the use of sociolinguistic interview data for purposes of disclosing change in that we must be certain that we use only data taken from like passages in the interviews when we

generalize. Otherwise, we may be barking up the wrong tree since *synchronic variation between situations* has not been taken into consideration.

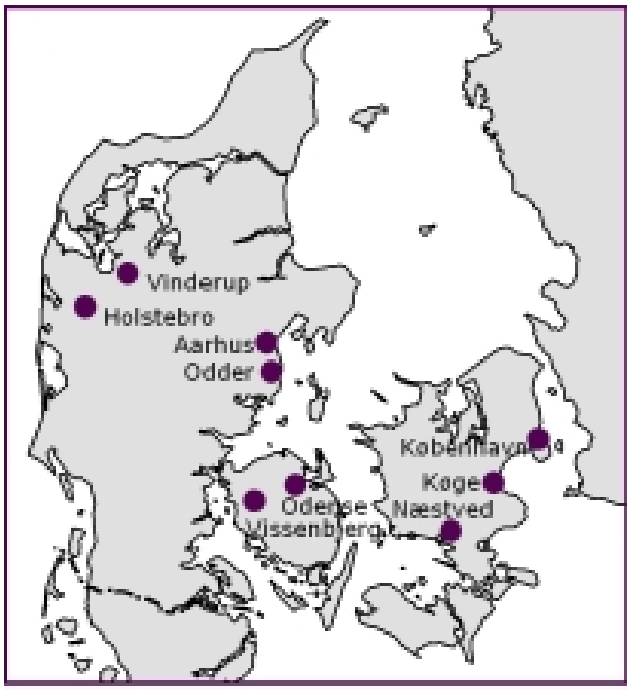
On comparability:
A study of language change in real time involves (at least) two studies, an earlier study (S1) and a recent one (S2). But how can we be sure that the studies are at all comparable? If the answer is a simple: "Because we use the sociolinguistic interview as the central means of data collection", we are begging the question: How can we be sure that the sociolinguistic interview is the same speech event, then and now? The empirical question is this: Given S1 and S2, both of them involving sociolinguistic interviews, can we find any differences in their internal structure that can be significantly related to the time of recording?
Taking the cue from the study of style shifting we shall have to make internal differentiations anyway and this may actually be a solution to the comparability question. (If, that is, there is any solution at all; I, for one, have much sympathy for a radical historicism which stipulates that any event in time is unique. The only trouble with this view is that it is unbearably skeptic and thus will, literally speaking, mean an end to all comparative research.)

The LANCHART project is designed to answer such questions as those posed above. We intend to repeat previous recordings from a total of 7 different projects carried out between 1973 and 2001 and covering six different sites all over Denmark. cf. fig 1



The Sites of Sociolinguistic Studies to be Replicated by LANCHART
Fig 1

The sites differ in type as well as linguistically, although by any standard Denmark ranks as a very dialect leveled speech community with only few, but readily perceived, geographic and social differences. To bolster our analyses by other types of data we have carried out language attitude studies at all the six sites as well, cf fig.2.



The Sites of Sociolinguistic Studies to be Replicated by LANCHART centre with reference points for the language attitude studies
Fig. 2

The Language attitude studies (Kristiansen 2007) taken as a whole present a picture of a speech community with sharply delineated conscious as well as non conscious norms which are strikingly similar across geographical sites and social classes.

The Vinderup data set

Taking the sites from West to East and in more or less chronological order, we have first the Vinderup studies. In 1973, Kjeld Kristensen, now at the Swedish Danish dictionary in preparation at the DSL, carried out two major spoken language investigations. In the first (in the LANCHART project called Vinderup 1), a sample of age stratified Danes were recorded in very short interviews (up to ten minutes) in order to ascertain how far the dialect leveling process had progressed in this speech community rather far from the major centers of Denmark, i.e. Århus and Copenhagen, cf. the map (fig.2).

In the LANCHART project, Vinderup 1 will be used to establish the point of departure in the beginning 1970ies for the dialect leveling process which we may follow at this particular site through three generations. The reason for this is that in 1978 Kjeld Kristensen returned to Vinderup to carry out a new project (Vinderup 2), based on the ideas of Mats Thelander and John Gumperz that instead of using the interview as the only medium of data collection, the contrast between formal and casual speech would be a consequence of the type of speech event sampled. Kjeld Kristensen consequently sampled the speech of 24 pupils of the eighth grade in two situations, formal interviews and group sessions. This study, called Vinderup 2, has been repeated in 2006 using only sociolinguistic interviews by Malene Monka and Signe Wedel Schøning at the LANCHART centre (Vinderup 2 new). They managed to find and interview 19 of the original 24 informants, thus creating a data set of 19 old recordings (in two situations) and 19 new interviews. To complete the picture, the same field workers gathered data from a total of 33 9th graders in

Vinderup so that the whole data set consists of old and new recordings with 18 informants and new recordings of 33 individuals who may be recorded by a new project in 10 or 15 year's time.

The Vinderup data set	1973	1978	2006	LANCHART Total
Vinderup 1	113			
Vinderup 2		24	19	43
Vinderup 3			33	33
Sum total	113	24	52	76

Table 1 The Vinderup data set (numbers represent recordings)

The Odder data set

In Odder, an age and geographically (town vs. country) stratified sample of 82 informants interviewed by the Danish dialectologists Bent Jul Nielsen and Magda Nyberg make up the first and oldest data set. Thisa data set is from 1986-87 when most of the data for the LANCHART project were gathered. 53 of the original Odder informants were re-recorded by Malene Monka and Signe Wedel Schøning of the LANCHART Centre in 2005 and of these we have created a focus group of 24 persons. The informants picked for the focus group were all between 25 and 40 years of age at the time of recording and are evenly distributed as to the speaker variables of social class (Working Class, WC, and Middle Class, MC) and gender (6 persons in each cell). This data set then consists of 53 persons (with a focus group of 24 persons) recorded twice in sociolinguistic interviews. Only the focus group interviews have so far been transcribed and analyzed, but the remaining interview data will be used for strategic projects focusing on particular variables or problems. To supplement the Odder 1 group we have created and Odder 2 group consisting of informants born between 1978 and 1987 (24 informants) and an Odder 3 group consisting of 33 ninth graders and their age mates (born between 1987 and 1990). These two groups have so far only been interviewed once, viz. by the LANCHART project.

The Odder data set	1986-87	2005-06	LANCHART total
Odder 1	82	53 (24)	106 (48)
Odder 2		24	24
Odder 3		33	33
Sum total	82	110 (81)	163 (105)

Table 2 The Odder data set

The Vissenbjerg data set

The original Vissenbjerg study was carried out by Inge Lise Pedersen in 1982-83. She interviewed 54 informants from her own home town in order to study the influence of the local and the national standard on this small town. Vissenbjerg is located quite close to the regional center of Odense (cf. the map, fig.2). 12 of the original 54 informants were re-interviewed by Henriette Simonsen in 1996.

The LANCHART group led by Tore Kristiansen has gathered new language attitude data from Vissenbjerg, as well as from the other sites, but we have not re-interviewed the informants. This data set then consists of 12 informants in old and new recordings.

The Vissenbjerg data set	1982-83	1999-2000	LANCHART total

Vissenbjerg	54	12	66 (24)
Sum total	54	12	66 (24)

Table 3 The Vissenbjerg data set

The Næstved data sets

Næstved must rank as the most studied town in Denmark sociolinguistically speaking, since two original and very far reaching studies were carried out in the period 1986 to 1989. One was directed by Tore Kristiansen (subsequently named Næstved I) and the other by Jens Normann Jørgensen and Kjeld Kristensen (consequently Næstved II).

The Næstved I data set

In Næstved I which focused on explicit and implicit language attitudes, Tore Kristiansen himself interviewed 48 adults, 39 youngsters enrolled in further education (9 years plus) and 36 kids attending compulsory education. Some of the kids were even interviewed twice. Of this impressive number of informants, the LANCHART center's Dorte Greisgaard Larsen and Lisbeth Bjerregaard in 2005-06 re-interviewed 34 adults (and duly created a focus group of 24 comparable to that of the Odder project).In 2006-07 a group consisting of Andreas Stæhr, Astrid Ag and Rikke Vivian Lange interviewed 19 youngsters and 18 of the original kids. The data set from Næstved I thus consists of 34 adults (with a focus group of 24), 19 original youngsters and 18 original kids interviewed twice in sociolinguistic interviews. Finally, the same group of field workers collected both 32 single person interviews and 8 group discussions with the same participants at four different schools in the Næstved region thus creating the data set of Næstved 3.

The Næstved I data set	1986-89	1992	2005-07	LANCHART total
Adults	48		34 (24)	82 (72)
Youngsters	39		19	58 (38)
Kids	36	12	18	66 (48)
Næstved 3			40 (32)	40 (32)
Sum total	123	12	101 (93)	246 (190)

Table 4 The Næstved I data set

The Næstved II data set

Næstved II was based on two ideas. One was that the style continuum could best be studied using the method that Kjeld Kristensen had used in Vinderup, viz. comparatively short formal interviews and rather longer group sessions without any interviewer present. The other was that a longitudinal study of youngsters beginning school and re-recorded once every year they were at the schoolswould reveal how young people accommodate their speech during successive stages. The LANCHART Centre plans to re-record as many of the original informants as we can in 2007. The original study comprised more than 60 students recorded most of them three times.

The Køge data set

The Køge study is internationally well known as a study of bilingual school children having Turkish and Danish as their first and second language respectively. It was established by a group of researchers including Jens Normann Jørgensen, Anne Holmen and Jørgen Gimbel. As of now, the Køge study is being documented thoroughly by Jens Normann Jørgensen (fthc.), and he and Janus Møller are in charge of the LANCHART follow up study.

The Køge study originally followed a core group of 12 kids from their first day at school and until they left school at 15 years. They were recorded in various types of sessions, sometimes in sessions with bilinguals and sometimes together with their Danish-only speaking age and class mates. A comparative study of the linguistic development of Turkish school children in Turkey (Eskisehir) is still ongoing.
In the follow up study, 11 of the original 12 core informants have been recorded in a number of situations: interviews in Turkish, (performed by Mediha Can: 11), 7 semi-controlled group conversations, 5 non-controlled group conversations all of them collected by Janus Møller. At the time of writing, Janus Møller estimates that 11 interviews and 3 group conversations have to be collected before the repetition is complete.
As with the other projects, we have established a Køge 3 group consisting of 8th graders. Janus Møller and Matthias Reichert and Louise Yung Nielsen have collected a total of 11 interviews in Turkish, 21 interviews in Danish and 10 semi-controlled group conversations. The Køge 3 data set is complete.

The Copenhagen data sets
In 1986-88, field workers at the so-called Copenhagen Project in Urban Sociolinguistics (BySoc) collected data from a total of 83 informants with a focus on individuals who were at the time of recording between 25 and 40. The informants were contrasted as to social class and most, but not all of them, had been born and raised in the historical neighbourhood of Nyboder, a characteristic ancient row of connected apartment houses designed and reserved for workers and officers affiliated with the Danish navy. The focus group of the new study of the Nyboder informants again was a group of 24 informants, evenly distributed among the cells containing two social classes and two genders. They were all except two re-interviewed by Janus Møller, the remaning two being interviewed by the present author.
The Bysoc corpus incidentally is probably more well known already than any of the other data sets in the LANCHART corpus due to the work by Peter Juel Henrichsen. He created the net based Corpus BySoc which was a modified version of the original transcriptions. The transcription practice was made uniform throughout, evening out any inconsistencies of transcription of the originals and these new transcriptions were subsequently publ ished on the net as the BySoc corpus (Henrichsen).
The Nyboder study, however, included a group of younger informants between 15 and 24 at the time of recording. From this group of 20 members of the BySoc 2, we have succeeded in collecting new interviews with 19, although two of them since they were not born and raised in Nyboder were not included in the original study's age group I but relegated to the so-called control group. Since no differences were found which were peculiar to the Nyboder neighbourhood, these two informants were included in the new study. BySoc 2 data were collected by Matthias Reichert Nielsen and Louise Yung Nielsen.
The Bysoc data group then consists of 24 informants interviewed twice, i.e. in 1986-88 and 2005-06. Some 5 of them also participate in group conversations with other informants. Furthermore 19 informants of the BySoc 2 group have also been interviewed twice, i.e. in 1986-88 and 2006-07. For a group of eighth or ninth graders from Copenhagen comparable to the Vinderup 3, Odder 3 and Næstved 3 data sets, we plan to use interviews already recorded and transcribed by Marie Maegaard.
Tore Kristiansen and his associates have collected new data on language attitudes from the Copenhagen region. These data are remarkably uniform and remarkably consistent with data from all the other regions (Kristiansen 2007)

The BySoc Copenhagen data set	1986-88	2006-07	LANCHART total
BySoc 1 (adults in 1986)	63	24	87 (48)
BySoc 2 (youngsters in 1986)	20	19	39 (38)
BySoc 3 (9th graders)		??	??
Sum total	83	43	126 (86)

Table 5 The Copenhagen data set

Finally, the DASVA project from Copenhagen 2001 studied the possible influence of the bridge between Denmark and Sweden on the neighbouring speech communities. The project studied both language attitudes to various types of Danish and Dano-Swedish as well as carried through 16 interviews with Danes between the age of 20 and 50. The data set has been transcribed using the Childes conventions and will be used to control the possible spread of rapid changes in the Copenhagen speech community since the data stem from only 6 years ago.

Current status

Most of the interviews mentioned as belonging to focus groups or supplementary groups have been transcribed. Some of them have been transcribed using the Childes conventions modified so at to minimize deviations from standard Danish orthography. As of late 2006, the LANCHART Centre changed to using the Transcriber program and developed a new manual of conventions. This has created a need for converting data from one format to the other and Peter Lind, the IT-chief of the project, has duly developed conversion programs from the various formats into Praat which is used for the analysis throughout. If I may allow myself a brief complaint here: It is really a great problem that we do not have one and only one program which is suited to the needs of the practicing sociolinguist viz. transcriptions, print outs and import to text programs, analysis and export to statistical packages.

The analyses of the LANCHART project

In the LANCHART project, it is our firm belief that no linguistic level should be studied in isolation. Consequently, we have worked with until now four different types of analysis:

- the language attitude study
- the discourse context analysis
- the sociophonetic analysis
- the grammatical analysis

Each of the types of analysis poses its own problems for the corpus design.

As mentioned above, *language attitude surveys* have been carried out at all the various sites using the methods developed by Tore Kristiansen and his associates during a period of 20 years (Kristiansen et al. 2005). Tore Kristiansen distinguishes explicit or conscious language attitudes from implicit or non-conscious. Explicit language attitudes may be collected by asking question about how various more or less well known varieties of Danish are evaluated. In all the sociolinguistic interviews carried out by the LANCHART Centre such explicit language attitudes are elicited by playing three or four different 'voices' for the informant while s/he is asked to place him- or herself with respect to the voices. The voices are so to speak being used as reference points for a geometric space of language norms.

Non-conscious or implicit language attitudes are collected by playing carefully prepared tapes including voices representing the three varieties relevant at the particular site, namely two Copenhagen-based Standard varieties, 'conservative' and 'modern', and the 'locally coloured'

Standard variety that is spoken by the local youth. In sum, we have explicit attitudes for all interviewees, and in addition we have both explicit and implicit attitudes from a representative sample of youngsters in each speech community (except Køge), taken from the compulsory school's final year.

The discourse context analysis was developed as a way to meet the needs of comparison. As you can see from the above, the original projects used at least three different methodologies, one relying on standard dialectological interviewing (the original Odder study), another putting the idea of operationalizing style as speech events to good use (the original Vinderup 2 study, the Næstved II study, both the original and the new one, and partly the Køge study, both the original and the new one), and finally the BySoc study, the Næstved I study, the Vissenbjerg study and the DASVA study used traditional Labovian techniques. Informants were also selected according to different criteria and there is an essential difference in the frequency and centrality of group conversations or double person interviews among the original studies: In the Vinderup and Næstved II studies as well as in the new Køge study, the interviewer is never present at group conversations - although he is not far away – in the event of equipment breaking down or any other such disaster.
This means that the data material is simply too various to be comparable in any simple sense. It has to be analyzed in at least foru dimensions. We have thus developed a coding manual which calls for an analysis of the *type of speech event*:

- single person interview with informant known to interviewer
- single person interview with informant unknown to interviewer
- group interview with informants known to interviewer
- group interview with informants unknown to interviewer
- group conversation without any interviewer present

Next we code for *activity types*:

- interview about informant's social background
- conversation
- conversation with a non-participant present or not present (e.g. calling on the telephone)
- elicited speech
- language attitude study
- informant's signing a declaration of permission to use the material for research purposes dependent on confidentiality

As you can see from this list, the categories are pragmatic. These activities are the activities which actually occur. Sometimes we ourselves make them happen (the signing of the permission is an obvious example), sometimes it is rather the opposite (conversation with a non-participant), although precisely these episodes may give us a glimpse of the reality of the observer's paradox.

The coding for *macro speech acts* is intended to catch the essence of what the transactions are about:

- exchange of knowledge
- exchange of attitudes
- exchange of emotions
- speech accompanying action
- fiction or phantasizing, repeating another person's speech or reading written sentences aloud

The last category was among other things invented to account for the invention of characters and their speech which occurred during one of the group conversations.

A classic in this connection is *the interaction structure*. The basic categories are those of initiative and response, and in the codings, we distinguish between short and long initiatives and responses:

- interviewer initiative with short response by informant
- interviewer initiative with long response by informant
- long interviewer initiative with short informant response
- no discernible structure
- informant initiative with response by interviewer
- fight for the floor
- informant initiative with response by other informant
- monologue
- residual

Here, it is even more obvious that the development of categories has been informed by the present data, since the informant response with response by another informant is dependent on there being another informant present, which is only the case when two informants are interviewed together (most often married couples) or a family group session takes place.

The first four coding categories described above have all of them been of the type where all that there is in the data has to belong to one or another of the categories. The next two coding levels are of the type where we only code passages where relevant. Let me here briefly mention that coding is solely performed on transcripts in order to avoid circularity: We might be influenced by the tapes in the discourse context coding, but we want to use the discourse context codes to control or elucidate sociophonetic variation.
The two remaining categories are genre and enunciation.

Codes distinguishing different genres are the following:

- narratives
- specific account
- general account
- soap box
- gossip
- confidences
- reflections
- jokes

Obviously, more genres would be needed to account for what goes on in counseling, doctor patient communication and so on, but these genres have all played a major role in the musings about style, they all of them actually occur during sociolinguistic interviews although they vary drastically in frequency, hence we have included them.

Enunciation is relevant for the coders because we want to isolate any passage where the speaker tries to imitate another speaker, makes illustrative noises, uses words *materialiter* or directly quotes someone, or finally reads aloud. In all these cases a shift of level of enunciation takes place, and we want to take note of that - if only because it may be essential to be aware that the phonetic variants used in such passages do not necessarily ‘belong’ to the speaker but rather to the person imitated.

The discourse context coding will be used to delimit comparable passages in the various data sets so that we take every precaution only to compare likes and to control what the likeness consists in.

This is necessary for the phonetic analysis to be operational, since it is evidently impossible to code all variants used from start to finish in more than 200 interviews which some of them are more than two hours long. Thus, a delimitation of passages to code for phonetic variation has to take place, and it has to be principled. This is the *raison d'être of* the discourse context coding.

The phonetic analysis

The phonetic analysis has so far been restricted to a start-to-finish study of a sub-set of 19 different files selected from the total corpus. The files were selected in order to maximize the differences in the total data set of the LANCHART corpus so that we could make a principled decision as to which passages were comparable and thus should be coded in the various data sets.. The variables studied have been

- the AN variable – (a) before alveolars and nothing
- the AM variable – (a) before labials and dorsals
- the AJ variable – the (aj) diphthong
- the ÆNG variable – the raising of (æ) to (e) before velar nasals
- the RÆ/RA variable – the lowering of short (æ) and (a) after r
- the RU/RO variable – the lowering of short and long (u) to (o) after r

The 18 recordings which have all of them been analyzed for discourse context and for grammatical features as well (cf. below) have been analyzed using the extremely time consuming process of listening to all tokens occurring throughout the interview or group conversation and categorizing every instance (upwards of 7000 tokens for one (long) conversation). All codings have been double checked for inter-coder agreement. The result is a data set with codes for all instances of every phonetic variable mentioned above - some of which are in fact rather infrequent. Currently, we are trying to conclude what type of discourse context, judging by this exploratory data set, is frequent enough in as many of the data sets as possible and in addition contains enough instances of the infrequent phonetic variables to be worth analyzing in the rest of the total corpus. Additionally, we may make decisions on the basis of the discourse context codes and the socio-phonetic analysis as to which passages to analyze in the various subsets consisting of old and new data from the seven sites.

The grammatical analysis

Three main variables have been studied in detail by post doc Torben Juel Jensen and his group of students. The first one is the generic use of various pronouns, the second is the reflexive system of third person pronouns and the third is the variable of main clause word order in dependent sentences.

Generic pronouns: In Danish, as in colloquial Canadian French, the second person pronoun, in this case Danish 'du', is spreading as a generic pronoun substituting for, or even taking over the role of, the generic third person pronoun 'man'. The process has been tied to the spread of English (Juel Jensen fthc.), but the origin is uncertain and could very well be local in that this particular use is known from Nordic dialects and has a long history before spoken language data corpora became available.

The use of *reflexive versus non-reflexive forms of third person pronouns* in spoken Danish has never been studied quantitatively until the present study. The norm of standard Danish as regards reflexively used pronouns may very well primarily be maintained via the educational system where the norm regarding the use of reflexive versus non-reflexive pronouns is explicitly taught, and deviation from the norm is diligently corrected by many Danish teachers. Even though the general impression seems to be that the non-reflexive forms are gaining ground at the expense of the reflexive forms when used co-referentially with singular subjects, and vice versa when used co-

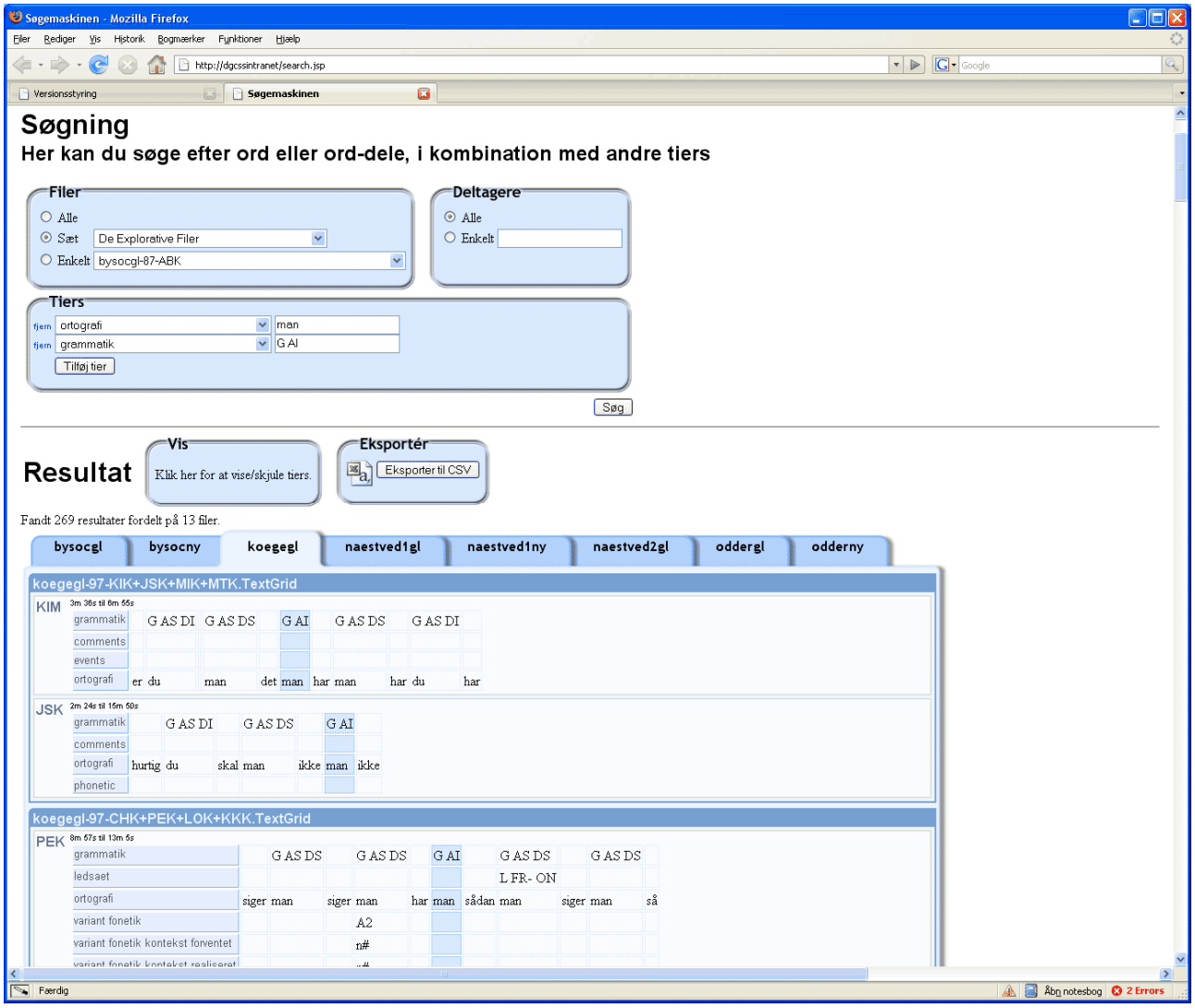
referentially with plural subjects, it may very well be the case that there has always been wide-spread variation in reflexive constructions, also among speakers of standard Danish/Copenhagen sociolects.
The third variable for which Torben Juel Jensen developed a coding manual is the use of main clause word order in dependent sentences. This variable has been studied in detail by Gregersen and Pedersen as well as by a number of formal grammarians, prominent among them Sten Vikner (Gregersen and Pedersen 2000).

The grammatical coding is to a certain extent dependent on the part-of-speech coding which was carried out for the LANCHART project by Peter Juel Henrichsen using the PAROLE tag set.
The plan is to have all files transcribed and converted to the same format at the end of 2007 part-of-speech tagged by Peter Juel Henrichsen so that all the material has the same structure. As of now, only the Odder 1 and Næstved I focus group interviews have been part-of-speech tagged.

The search engine and the file director
The recordings are transcribed using the program Transcriber. They are then converted to Praat and all codes use the Praat format. The process with the proper treatment of so many files from so many different projects is very complicated and Peter Lind has consequently developed a file director. system which ensures that a file is locked to the user, once the user has been licensed to engage with it and that no other person can break in to 'steal' the file. Furthermore, the file director system makes it possible to manage the process in that files may be loaded into the system but kept there invisible for the users until they are licensed to be transcribed or analyzed.
Developed by Peter Lind, the search engine operates on coded Praat files. It is possible to search the data for specific informants, using the informant data base which includes all persons who have an informant code. This may be the case also for participants in conversations who simply happened to be there or to visit while the interview was going on. In a particular case, which I have myself experienced,. the informant's daughter came by during the interview to take a bath. Her bathroom was out of function and so she used her father's instead. When she had finished her bath, I went to the toilet. The recording, however, continued while I was away. Looking at the transcripts of the passage when father and daughter were alone, one of the Discourse Context coders, Astrid Ag, drew my attention to the fact that the informant consistently used more swear words than he did in conversation with me. Again, a vivid demonstration of the observer's paradox.

It is further possible to search for any kind of code and combine the codes from all levels of analysis. Thus we may want to know whether there is any discernible difference between the a-variables (AN, AM and AJ) used in narratives compared to that used in interview passages which have been coded as concerned with speaker background information. The output will be a .csv file which without any problem may be imported by the SAS statistics program or by Excel and thus be manipulated further in those programs.

I have borrowed one of Peter Lind's screen dumps demonstrating the interface:



Under the headline ’Filer’ you may specify what file in the corpus you want to work with. The options are: all files stored in the server, a subset or a single file. You may further pick one or several of the participants in the box named ‘Deltagere’. Options here are either all of them or a single informant, in which case you will have to specify which one.
In the box called ‘tiers’, you may choose between the various coding tiers, viz. orthography, grammar, discourse context or the various phonetic codes. In the ‘result’ section, you may either view the result which is listed in a modified Praat format (this is essential in order to check the result before it is exported for statistical operations) or go directly to the Export function. As mentioned above, the export format is .csv.
Finally, the various projects have each of them got their own file, so that you may easily switch between the data sets when searching.

Conclusion

The most original feature of the project, and now also the search engine, is the possibility of looking at combinations of codes when searching through single files or subsets or even the total corpus, using the whole gamut of linguistic levels to create new insights. Further levels to be included in the future are lexical semantics which may conveniently sue the orthography tier as the point of departure to study changes in the choice of lexical items in real time.

When describing a corpus such as the LANCHART corpus of spoken Danish in real time, one may use various units to give the reader an impression of the versatility and character of the corpus. First of all, we may look at how many informants are in the corpus. The answer is that between 680 and 730 persons have been involved in recordings. The next, and probably more important, measure is how many files, viz. unique combinations of informant and recording are represented. This is more or less the method used in the tables above. Finally, we may use the traditional measures of how many hours of speech or how many words or bits we have in the corpus. The answer to this obviously relevant question is simply: I do not know, but plenty for us to work with!

References

Gregersen, Frans and Inge Lise Pedersen: *A la Recherche du* Word Order not quite *perdu*, in Herring et al. (eds.): *Textual Parameters in Older Languages*, Current Issues in Linguistic Theory 195, p. 393-431, John Benjamins, Amsterdam 2000

Henrichsen, Peter Juel: Corpus BySoc http://www.id.cbs.dk/~pjuel/cgi-bin/BySoc_ID/index.cgi

Henrichsen, Peter Juel: *Talesprog Med Netstrømper - Internet-adgang til et stort dansk talesprogskorpus"*; 93pp; Cph Univ.: Instrumentalis 12/1998

Jensen, Torben Juel: ms. on grammatical changes, submitted for publication in The XXth meeting on the Grammar of Danish

Normann Jørgensen, Jens: The Køge project, fthc.

Kristiansen, Tore: Sprogholdningsundersøgelserne. Kortfattet rapport om oplæg og hovedresultater, http://dgcss.hum.ku.dk/upload/application/pdf/f51d6748/Sprogholdningsundersøgelserne%20samlet.pdf, ms. 19. april 2007

Kristiansen, Tore et al. (eds.): Subjective Processes in Language Variation and Change, *Acta Linguistica Hafniensia* vol 37, Reitzel Copenhagen 2005

Labov, William: *The Social Stratification of English in New York City*. Center for Applied Linguistics, Washington D.C. 1966, second edition 2006

Labov, William: Field Methods of the Project on Linguistic Change and Variation, Baugh and Sherzer (eds.): *Language in Use. Readings in Sociolinguistics*, pp.28-66, Prentice Hall, Englewood Cliffs 1984

Meyerhoff, Miriam: *Introducing Sociolinguistics*. Routledge, London 2006

Milroy, Lesley and Matthew Gordon: *Sociolinguistics. Method and Interpretation*. Blackwell Oxford 2003

Sankoff, Gillian: Cross-sectional and longitudinal studies in sociolinguistics, Ammon et al. (eds.): *Sociolinguistics, an International Handbook of the Science of Language and Society*, vol. 1, de Gruyter, Berlin 2005