Peeking Into the Danish Living Room
Internet access
to a large speech corpus

Peter Juel Henrichsen

Dept. of Danish Dialectology and
Dept. of General and Applied Linguistics (IAAS)
Univ. of Copenhagen, Njalsgade 80, DK-2300 S, Denmark
pjuel@cphling.dk

1. Introduction

Our newly opened Internet site offers a view to a $>10^6$ word corpus of informal Danish conversations. The corpus and the search engine situated at IAAS can now be reached and used as easily as any homepage on the World Wide Web, offering a tool for serious investigations into informal speech.

After a few introductory remarks, we shall present the corpus and the search engine as seen from the user's point of view, following up the presentation with a few example queries. In conclusion, some reflections on the possibilities that the Internet has to offer in utilisation, maintenance, and control of large corpora of semi-confidential data.

The new site may be reached directly at the URL **http://phoneme.cphling.dk/BySoc**, or else via the IAAS home page at **http://www.cphling.dk**

1.1 Why study informal speech?

Ordinary people are common. Most of the day, they talk casually, taking part in informal conversations. So, by far the largest part of the language national product must be plain and simple vernacular. Moreover, most children acquire their mother tongue exposed to this style only, arguably making it the most essential part as well.

Still, the syntax and semantics of informal speech have hardly been studied at all with the exact means of modern formal grammar. Why?

Firstly, because informal speech is *irregular*, seen from a traditional syntactic point of view, making it much more recalcitrant to work with than educated written style. Secondly, because it is hard to get access to large, reliable samples of genuine everyday conversation; the use of bugs is of course unethical, and perhaps illegal, making this abundant source of language rather elusive as a scientific object (which may well seem a bit paradoxical). Finally, because generative linguistics have turned these two obstacles into a claim that informal speech is theoretically uninteresting.

This paper suggests a solution to one of the problems.

2. The corpus

The searchable corpus consists of transcriptions of approximately 80 conversations[1]. The participants are all native Danes, most of them born and raised in Nyboder in central Copenhagen. In each case, the linguist who participated in the conversation took great care to make the situation as comfortable and familiar for the informant as possible. Most of the conversations thus took place in private homes, with the linguist as a so-called ‘intimate stranger’, rather than a TV-style interviewer. The strategy was the following: Prior to his first meeting with the informant, the linguist would try to get acquainted with a neighbour or such, enabling him to present himself to the informant as ‘a friend of your friend’. “Normally, people do not confide in a stranger, but some have done so when they have met *an intimate stranger*” (Gregersen & al (91) p.53, following Milroy & al (77)). After a while, many informants reach a relaxed, or even confidential state. This state is of course desirable, since it is exactly then that we get the informal style free of conscious control. For this reason, the sessions recorded were rather long, almost always more than an hour, in some cases more than three hours.

The recordings were transcribed using the convention Dansk Standard 2, as defined by The Danish Language Council (Dansk Sprognævn), a basically orthographic code, enriched with special symbols. In addition to Dansk Standard 2, a so-called *score format* was designed, showing the onset of the utterances relative to each other.

Transcription sample

```
------------------------------------------------------------------------------------------
1>                          ja ja                                                 nej~
2>                          nå nå (ler) # nå I startede i de små lejligheder ik'       nå
A>ja vi starter helt forfra (ler)
------------------------------------------------------------------------------------------
1>
2>£ det troede jeg ?(ellers)?                                          ja ja

A>                 ja fordi der der er dem der er mindre £ der er dem der er på halvandet
------------------------------------------------------------------------------------------
1>på~         atten kvadratmeter         ja~                    sådan en fik vi tilbudt £
2>                                           ja det var hvad vi kunne få

A>værelse ik'                    ja~  £

------------------------------------------------------------------------------------------
```

[1] The searchable corpus covers about two thirds of the recordings made in connection with the project *Urban Sociolinguistics* (Projekt Bysociolingvistik), carried out in the late eighties by a group of Danish sociolinguists rooted in Institute of Danish Dialectology and Institute of Nordic Philology at the University of Copenhagen. An important purpose of the project was to study Danish vernacular as a function of age, sex, and social class. However, according to leading member Frans Gregersen, the group was very anxious not to have this particular purpose influence the data collection. The interviewers were thus instructed to let the informants choose as well topic as style, not urging them into narration, say (pers.comm.). The background and results of Project Bysociolingvistik are reported in Gregersen & al (91).

Special symbols: A> refers to an interviewer; 1>, 2> ,... to informants; ~ stands for 'tøven' (hesitation); £ for 'pause' (pause); # for 'pause med vejrtrækning' (breathing); (...) for 'transskriptørkommentar' (transcriber's comment) ?(...)? for 'usikker transskription' (transcription uncertain)

The searchable corpus (*BySoc* from here on) consists of 1.4 million words, distributed over some 35,000 word forms. The corpus has recently been thoroughly proof read and converted into a well-defined data structure (cf. Henrichsen (97)). *BySoc* has no grammatical annotations.

3. The search engine

The central window in the *BySoc* site consists of a number of frames, among which the most important are the frames **Søgeområde** (Search Area), **Søgeprofil** (Search Profile), and **Resultat** (Result). Using the buttons and tables in these frames, the user can specify

- A search area (a subpart of the corpus)
- A search profile (a string of letters and special symbols)
- An output format (*count*, *list of occurrences*, or *text samples*)

3.1 Search Area

The search domain can be cut along the following dimensions (among others).

Speaker:	sex, age, role (linguist/informant), social group (working/middle class), name, id
Conversation:	number of participants, time slice (e.g. 'exclude the first ten minutes')
Style:	language (*Danish/foreign*), enunciator(*self/quote*),origin(*transcription/comment*)
Transcription:	transcriber, confidence (*reliable/uncertain*), version (*main/secondary/encrypted*)

3.2 Search Profile

A simple orthographic word will do as a search profile, while in general any *regular expression* is allowed. Our dialect of RE is as in the programming language *Perl* (Wall (91)), enriched with special symbols for 'space character' (_) 'letter' (¤), 'vowel' (½), and 'consonant' (§). Some special symbols of RE: *x*? matches zero or one instances of *x*; *x*{*n*,*m*} matches *n*-*m* instances of *x* in sequence (*m* may be omitted if there is no upper bound); *x*+ is equivalent to *x*{1,}; *x** equivalent to *x*{0,}.

Example profiles

Search profile	**matches**	**does not match**
Nyboder_Skole	Nyboder Skole	NyboderSkole
[Nn]y.ode(r|ns)	Nyboder nymodens ny oder	yboder nyodens ny ode
§ybod?er	Nyboder byboer	Ayboder bybodder
(ja)*da	da jajajada	jjjda jaada
¤{1,2}y§od+er	Nyboder blylodder	yboder polyboder Nyboer
¤+	Nyboder slikasparges	Nyboder Skole slik- asparges

Some linguistically relevant profiles are demonstrated in sections 0 and 0 below.

3.3 Result format

Different queries call for different kinds of output from the search engine. The user can choose output format by clicking one or more buttons in the table 'Resultatets form' in the frame **Resultat**. There are three families of output formats available.

- Count number
- Annotated word list, sorted by frequency or alphabetically
- Text samples, in one or two dimensions (single speaker or full score)

The user may also choose to have the result sent as email.

3.4 Query check

While the user is working on his query, his dispositions are being checked for consistency by a background process. This process is controlled by a *JavaScript* program (cf. ref.), embedded in the HTML source code. The process is invisible to the user until he makes an illegal choice – such as excluding men *and* women from the search area, or putting restrictions on '$3' while his profile contains only *two* bracketed subparts. In such cases, the background process corrects the error, if possible, or otherwise resets the parameter in question to its default setting.

A simple example is the check procedure associated with the query parameter *sex*. The corresponding table in the frame **Søgeområde** (Search area) is generated by the following HTML code.

HTML sample

```
<a name=sex><i><u>Køn</u></i></a><br>
<select name=sex multiple onchange=notempty(this,2)>
  <option value="F" selected>kvinder
  <option value="M" selected>mænd
</select>
```

Each time the user clicks on one of the options, *male* or *female*, the function **notempty** is called.

JavaScript sample

```
function notempty(s,e) {
  var flag=false, i=-1;
  while (s.options[++i]) {
    if (s.options[i].selected) {flag=true; break}
  }
  if (!flag) {for (i=0; i<e; i++) {s.options[i].selected = true}}
}
```

If the user happens to deselect *all* options – in the case at hand: both sexes – the background process resets the table to its default setting [*male* **selected**, *female* **selected**]. Otherwise, the process stays invisible.

3.5 Submitting a query

When the user has finished his query, he clicks the **Start søgning** button (Start search). A JavaScript process then collects all parameter settings, encloses them in a virtual envelope, and submits them to the server. At the server side, a UNIX daemon unpacks the envelope, starts a Perl session, and feeds the search engine with the user input as % ENV variables (for technical details cf. Holtse and Henrichsen (*in preparation*)).

Before the result is returned to the client, certain proper names are encrypted, namely those that do not occur in Retskrivningsordbogen (the Standard Danish Orthographic Dictionary). This means that non-local names like *Grønland* and *Roskilde* are allowed, while *Delfingade* (a local street name) and *Margrete* (a personal name) are not. Of course, this does not exclude the possibility of misuse, but it does prevent the sincere user from being presented with confidential data unwillingly (in case of deliberate misuse, appropriate actions can be taken, since each user must sign a sincerity declaration to obtain a password).

The output of the search engine is then translated into an HTML document, enclosed in a virtual envelope, and sent via the Common Gateway Interface (CGI) to the client where it is interpreted and presented to the user.

4. Sophisticated queries

This section may be skipped by the busy reader.

4.1 Cascaded filters

One of the special features of our search engine is the *cascaded filter*. By enclosing certain subparts of a profile in angle brackets (< and >), one can refer to them individually. The first bracketed substring is then referred to as '$1', the second as '$2', and so on. By putting individual restrictions on these named subparts, one can construe sophisticated filters piecemeal.

Consider an example (this section may be skipped by the busy reader). Find adverbial clusters included in compound verb groups of the type 'har ... været' – clusters such as 'har sikkert nok været' and 'har jo heller ikke altid været'[2]. For practical reasons, we will restrict the goal to 2-6 word clusters.

As a first try, we define the profile **_har<_(¤+_){2,6}>været_** and specify that only $1 is to be registered, i.e. the 2-6 uninstantiated words between the angle brackets. Now the unbracketed parts of the profile serve as context constraints. This is not good enough yet, though. Along with the wanted adverbials, many subject NPs in detopicalized position are admitted: 'har du selv været...', 'nogen gange har min mor og far været...', etc. So we impose a restriction on $1 that it may not match **_(jeg|du|min|din|han|hun|nogen|alle|vi|de|I|det|den|der)_**. The words 'jeg', 'du', ... are the most frequent Danish pronouns in the nominative and indeed the most frequent subject NP elements in *BySoc*.

Now the profile works, but of course it misses out a lot of relevant adverbial clusters, namely all those contained in *other* participle constructions, such as 'har ... haft' og 'har ... kunnet'. So we soften up the profile: **_har<_(¤+_){2,6}>(haft|¤+et)_**, now allowing for a spectre of participle constructions such as 'har faktisk ikke haft en chance' and 'har nu ikke altid løbet så stærkt'. However, a new restriction is called for, since the **(haft|¤+et)_** part of the profile happens to match, not only participles, but also pronouns in the neuter, as in e.g. 'har et eller andet imod mig'. So we put **(haft|¤+et)_** in angle brackets and specify that $2 must not match **(det|noget|meget|andet)_**, the four by far most frequent $\text{PRO}_{\text{neuter}}$ in spoken Danish. These little adjustments can be added in a bottom-up fashion, simply by analysing the output of the search engine.

Now the filter performs quite well, making it reasonable to start drawing conclusions from the results delivered by the search engine. Still, there is plenty of room for improvements, on the one hand by extending the hunting ground with modal verbs ('kan'), new auxiliary verbs ('havde'), subordinate clauses ('at jeg ikke har haft den'), etc. – and on the other hand by sharpening the constraints.

Should we later want to focus on special adverbial clusters only – say, those containing negative polarity items, or 'ikke', or discourse particles like 'jo', 'vel' and 'sgu' – this can be done by simply adding more restrictions on the $-variables.

[2] In terms of the *feltskema*, Diderichsen's reknown description of the Danish sentence syntax, the adverbials in question belong to the *a*-slot of the nexus field (Diderichsen (46)).

4.2 Two-dimensional profiles

In informal conversation, the word changes rapidly and impulsively – the more so, the more typical the conversation. Hence, any analysis of the semantics and pragmatics of conversation without access to the distribution of utterances over speakers would be impaired, or even meaningless.

So of course such *BySoc* information should be accessible, and this calls for a new entry to the search engine. In order not to have the interface look like the dashboard of a Jumbo jet, we have chosen to add just one new convention: '%*n*'. As mentioned in the previous section, '$*n*' refers to the *n*th bracketed subpart of the profile. '%*n*', in contrast, refers to the same text position as '$*n*' (i.e. the same time slice), but *anyone but* the same speaker.

Consider an example. In Danish informal speech, the tag *ik'* usually triggers a positive response (we take positive responses to be *jo*, *ja*, and *mm*). Find all exceptions to this rule.

One solution would be to search for `_ik'_<_.{2}>`. This profile finds each utterance-final occurrence of *ik'*, thanks to the two adjacent spaces. The bracketed part of the profile then corresponds to the first 3 characters of the right context -- and when referred to with '%1', it applies to the full score, or rather: all speakers with the exception of the speaker saying *ik'*. Now we need only to impose the restriction on '%1' that it may not match `jo|ja|mm`.

Two of the *ik'* occurrences in question occur in the transcription sample in section 0.

5. Linguistic sessions (examples, not science!)

5.1 Trigrams, n-grams

Trigrams – groups of three adjacent words – are well-known creatures in corpus linguistics. *BySoc* trigrams may be found with the search profile `_(¤+_){3}` (easily modified to cover 4-grams etc.).[3]

The search engine reports 780,513 hits, the most frequently occurring trigrams being:

[3] Select the output format '*list*' (deselect '*text samples*'!). As queries of this kind put a considerable load on the server, you may have to split up the corpus. You could, for instance, cut it in three time slices: 0-49th minute of each interview, 50th-99th minute, and 100th-and-on, and search them individually.

Rank	Form	Count
#1	'det er jo'	1881
#2	'og sådan noget'	1589
#3	'men det er'	1340
#4	'i hvert fald'	1248
#5	'det er det'	1089
#6	'ja det er'	961
#7	'og det er'	858
#8	'det kan jeg'	800
#9	'det ved jeg'	783
#10	'jeg ved ikke'	781

5.2 Differences in word selection: men versus women

As mentioned, the corpus can be torn along various seams, including *sex* and *age*, providing for investigations into word selection. For instance, the frequency of a few pronouns, 'jeg', 'man', 'du', 'min', seem to vary quite a bit as a function of sex and age (search profile: `_(jeg|man|du|min)_`).

Frequency and rank	*'jeg'*	*'man'*	*'du'*	*'min'*
M, 14-25 years	3.8% (#2)	1.0% (#21)	0.63% (#31)	0.30% (#58)
M, 25+ years	2.7% (#5)	1.2% (#16)	0.80% (#27)	0.24% (#66)
F, 14-25 years	5.4% (#2)	0.71% (#29)	0.59% (#34)	0.36% (#45)
F, 25+ years	3.5% (#2)	0.85% (#25)	0.84% (#26)	0.34% (#49)

5.3 George Kingsley Zipf revisited

G. K. Zipf was one of the first to suggest general statistical hypotheses about language. The following has become known as 'Zipf's law' (Zipf's own formulation was a bit different, but equivalent).

In any corpus of considerable size, in any language, $\forall z \in \mathit{Word\_forms}: \mathit{Rank}(z) \times \mathit{Count}(z) = c,$ c constant

Zipf's claim does hold quite nicely in a variety of Danish text corpora, at least for words of rank >15.

Rank	Rank × Count / 1000		
	newspapers	*magazines*	*scientific journals*
#1	8	8	8
#2	13	13	15
#4	15	17	21
#8	25	27	31
#15	28	33	37
#30	22	27	25
#60	19	24	20
#120	22	24	19
#250	23	22	22
#500	25	24	23
#1000	26	24	25
#2000	27	24	24
#4000	25	23	23

Figures based on Maegaard &al (86)

Does informal spoken Danish show the same kind of regularity?
To answer this question, we compute a frequency list using the profile _¤+_, or, more efficiently[4]: _¤+.

Total: 1,348,083 hits.

Rank	Count	Wave length = Total / Count	Frequency	Accumulated frequency	Rank × Count /1000	Word form
#1	74191	18	5.50%	5.50%	74	'det'
#2	45779	29	3.40%	8.90%	92	'ja'
#3	41338	32	3.07%	12.0%	124	'og'
#4	39497	34	2.93%	14.9%	158	'jeg'
#5	39218	34	2.91%	17.8%	196	'er'
#6	36211	37	2.69%	20.5%	217	'så'
#7	32194	41	2.39%	22.9%	225	'der'
#8	25315	53	1.88%	24.8%	203	'ikke'
#15	17284	77	1.28%	35.0%	259	'ik"
#30	10337	130	0.77%	49.9%	310	'også'
#60	3467	388	0.26%	62.9%	208	'vil'
#120	1359	991	0,10%	72.6%	163	'ude'
#250	477	2826	0.035%	80.3%	119	'hvorfor'
#500	178	7573	0.013%	85.7%	89	'døren'
#1000	70	19258	0.005%	89.73%	70	'situation'
#2000	27	49929	0.002%	92.9%	54	'forstod'
#4000	10	134808	0.0007%	95.4%	40	'bøgerne'

[4] The engine has two search algorithms, one lazy and one meticulous. The lazy one is the faster, but it does not find overlapping hits, so it wont work with the profile _¤+_. On the other hand, it does work with _¤+, and this profile is (nearly) equivalent, since the + operator is 'greedy', always matching as many characters as possible.

As can be seen in the table, corpus *BySoc* does not follow Zipf's law. The distributional patterns of informal speech seem to be radically different from those of written texts. While this is hardly a controversial observation in itself, it does trigger a series of challenging questions: *How different*? Where? And why?

In conclusion, the reader is invited to try another Zipf-hypothesis on corpus *BySoc*.

> "The product of the number of words of a given occurrence, when multiplied by the square of their occurrences, remains constant for the great majority of the different words of the vocabulary in use, though not for those of highest frequency." (Zipf (36) p.41ff).

Visit **http://phoneme.cphling.dk/BySoc**, enter the profile _¤+, and select the output format *Zipf list*.

6. Concluding remarks

Using the Internet as a keyhole, there is little need for distributing copies of a corpus. A single copy residing at the server side is sufficient, for many purposes. This has some significant advantages.

Firstly, the researcher can document his results, simply by reporting the parameter settings of his central queries. Secondly, maintenance of the corpus is fast and efficient. Each user is a potential proof reader, and his reported corrections can be implemented and utilised in minutes or hours. Thirdly, confidential data can be encrypted, and if necessary, effectively barred within an instance. This improved efficiency of control can be foreseen to actually make the corpus *more freely* available, since the corpus administrator can now mete out access much more precisely – if necessary, even on an individual basis – and does not have to employ 'better too little than too much' kinds of precautions. The precision in control could also make the supplier of data feel more confident, as he can now be given (i) a qualified promise that no unauthorised party will gain access, and (ii) a detailed description of what authorised users will be allowed to see and do (one could even trade with a reluctant informant, offering the withdrawal of data in case, say, the private situation of the informant should change).

The client-server model offers *platform independence* on the client side, and *specialisation* on the server side. This may lead to a new and attractive situation: even the slowest of today's 486s or Macs is sufficient for the most powerful searches, as long as there is an Internet connection (and a willing corpus provider at the other end). The implicit anarchy and decentralisation of the Internet could thus strengthen the research community by counteracting a notorious technocratic and anti-individualistic barrier in the corpus based research. Via the Internet, everyone, including individual researchers and low-budget institutions, can have equal access to resources and methods. In the same vein, the technically insecure linguist can make as intricate queries as can the UNIX or Perl wizard.

Last but not least, no one is forced into buying a specific software product, thanks to the system independence of HTML and Java – an obvious democratic advantage, which the Department of Justice in Washington right now is struggling to preserve.

7. Acknowledgements

Thank you to the *BySoc* group for the enormous work they have done collecting this unique corpus. It must have been a battle. Thanks also to Dorte Haltrup Hansen who swept the battle ground for left-over misspellings and more.

The work reported in this paper had not been possible without the steady and unconditioned support from Department of Danish Dialectology, particularly Inge Lise Pedersen, for which I am very grateful.

8. References

Gregersen &al (1991) *The Copenhagen Study in Urban Sociolinguistics*, 1+2; Copenhagen: Reitzel.

Milroy, L. and J. Milroy (1977) *Speech and Context in an Urban Setting*; Belfast Working Papers in Language and Linguistics, vol. 2,1.

Henrichsen, P. J. (1997) *Talesprog med Ansigtsløftning*; IAAS, Univ. of Copenhagen: Instrumentalis 10/97 (in Danish).

Holtse, P.; P. J. Henrichsen (*in preparation*) *Multi-modal Corpus Presentation via Internet.*

Maegaard, B. &al (86) *Hyppige Ord i Danske Aviser, Ugeblade og Fagblade*; Gyldendal.

Wall, L. &al (1991) *Programming Perl*; O'Reilly.

Zipf, G.K. (1936) *The Psycho-biology of Language – Introduction to Dynamic Philology*; London: Routledge.