Annotation of spoken language data
Janne Bondi Johannessen and Fredrik Jørgensen

Introduction
Creating syntactic trees for spoken language data involves many challenges. The problems pertain to what the basic units are, whether there are units that should be disregarded, how to attach various units to each other, what to do with the fact that more than one speaker is involved in the discourse, what to do with intonation, etc. The results presented in this paper derive from the first phase of a larger tree-annotating effort in connection with the project NoTa (a corpus of spoken Norwegian). The first phase involves investigation of the spoken language data and a first syntactic tree-annotation using a hand-crafted phrase structure grammar. The future phases will look into the possibilities of using statistical methods to arrive at a full syntactic annotation. In this paper we shall first consider categorisations of the main features that characterise spoken language as presented in the literature on spoken Scandinavian languages, and then discuss to what extent, based on our experiences with applying the PSG-grammar to the NoTa corpus, these categories can be represented in syntactic trees.

Basic units
In order to create syntactic trees for spoken language units, it has to be clear what the units are. Any decision of course depends totally on one’s theoretical perspective. If one’s main interest is discourse functions, any occurrence of repetitions, interruptions, self-corrections, as well as appositions and even pauses, will be of major interest, maybe at the expense of annotation of hierarchical syntactic structures. If one’s perspective, on the other hand, is purely syntactic, one might choose to ignore the discourse functions mentioned above. Our view is that syntactic structure is important in spoken language as much as in written language. However, we are not willing to ignore the features that characterise spoken language. Therefore, we need to find ways of representing such features as well as the features known from written language.

Characteristics of spoken language
The characteristics of spoken Scandinavian language have been discussed and described by (among others) Teleman (1974), Hanssen et al. (1978), Wiggen (1986), Allwood (1998), Nygård (2004). The macro-syntagms recognised in a Norwegian investigation into the Oslo dialect, TAUS (Hanssen et al. 1978), are the following:

Wiggen (1986) divides these into two groups, interjection macro-syntagms (including interjections and vocatives) on the one hand, and sentences and sentence fragments on the other:
(1)
IMS – interjections:
ja, nei, oi
yes, no, oh
TMS - vocatives:
Du! Kehner!
You! Waiter!
AVBR - interruption:
folte du at du måtte forandre deg sjøl da eller eller h- holdt du på ... (NoTa)
you felt that you had to change yourself then or or d- did you ...
SyntK - syntactic correction:
ja hvis jeg flyt- la oss si at vi fl- jeg f- # bodde der fra # jeg gikk i åttende klasse
(NoTa)
yes if I move- let us say that we m – I m # lived there from # from I was in 8th grade
Utel - “missing material”:
em # har bodd ganske mange steder
ehm # have lived in quite many places
Anak - anacolut:
??Bogstadveien Hegdehaugsveien er det egentlig ganske forferdelig bortsett fra et par steder så er det # kun gutter i blå skjorte og # mørke bukser(NoTa)
The Bogstad road, The Hegdehaus road it is actually quite horrible except for a few places there are # only a few blokees in blue shirts and # dark trousers
Ans - start:
nei da # vi har det bra vi så (NoTa)
oh no # we have a good time so

Repetitions:
folte du at du måtte forandre deg sjøl da eller eller h- holdt du på ... (NoTa)
did you felt that you had to change yourself or or d- did you ...

Many of the above types can be regarded as interruptions or fragments with missing material. Telemann (1974:244ff) also points at other types of macro-syntags, viz. ones that have a specialised meaning, and which could structurally, but possibly not semantically, be seen as a well-defined part of a clause. Especially exclamations fall into this category. The examples below are Swedish, but they work equally well in Norwegian and possibly Danish. They are all well-formed spoken utterances, but differ from written norms in that they are only sentence fragments, subordinate clauses or lack a finite verb:

(2) (så) bra at du kunde komma!
how good that you could come
(3) som om jag skulle ha ljugit!
as if I should have lied
(4) att du inte kunde vara lite snällare!
that you couldn’t be a bit kinder
(5) han spelar fotboll!
he play football

How to relate syntax with spoken language as it has been exemplified above is a question that has to be answered. There are in principle three possibilities:

• Ignore all spoken language features (Meteer et al and Taylor 1995)
• Ignore some spoken language features and take others seriously (Nygård 2004)
• Take all spoken language features seriously (As done in the discourse literature)

The most drastic of these measures is the first one. Meteer et al and Taylor describe a three-
level annotation scheme, where the first level includes every little sigh and repair, while the second level involves a bracketing of all speech signals and even conjunctions and repetitions. The third level cleans the annotation away, and the text is left as simple main clauses only. The second method is suggested by Nygård (2004). She proposes that spoken language features can be related to either discourse level or structural level. Discourse level material includes interjections and vocatives, but also filler words (discourse particles) and some simple words and phrases. Material that belongs to the structural level fullfills one simple criterion: whether they can be paraphrased unambiguously. Consider (6):

(6) Mulig at de kommer. (The Big Brother Corpus)

What is left out from this sentence is the word string Det er ‘it is’. Thus, this sentence can be given a full structural analysis, including the elements that are left out. Nygård argues for the paraphrase criterion using several linguistic arguments. Although her arguments are convincing, they are based on theoretical linguistic argumentation and not language technology. Thus, although it is quite clear that some sentence fragment could be seen as part of some larger sentence structure, it would be easier for a human than for a computer program to guess what is lacking. One reason for this problem is the fact that a computer program will not know the full semantics of every word. Take (7)–(8) as an example. (7a) could only have the paraphrase (7b), with a formal subject, and (8a) only (8b), or others with a lexical subject. But structurally, in the eyes of the computer, the two a-sentences are the same, as long as the computer is unaware of the semantics of the individual words in the sentences.

(7) a. _jobbet en mann i hagen
    worked a man in the garden
b. det jobbet en mann i hagen
    there worked a man in the garden
c. * barnet jobbet en mann i hagen
    the child worked a man in the garden

(8) a. _jobbet en stund i hagen
    worked a while in the garden
b. jeg/mannen/barnet jobbet en stund i hagen
    I/the man/the child worked a while in the garden
c. * det jobbet en stund i hagen
    there worked a while in the garden

The third solution is to take all spoken language features seriously. This is what we will do. Miller and Weinert (1998:30) say that “it is far from evident that the language system of spoken English has sentences”. We disagree. We think that Nygård successfully argues for the fact that (most) sentence fragments are parts of sentence structures. Furthermore, it seems clear that spoken language is systematic w.r.t. “missing” parts. Wiggen (1986) shows that “missing” personal pronouns represent 48% of all missing words, while the formal subject pronoun det represents another 15%. We take these facts to support Nygård’s findings, and want to base our approach on the idea that spoken language data is structured.

While the second part of our project will look into spoken language in a fully systematic way, using both the TAUS texts and the NoTa (Norwegian Speech Corpus) material and statistical as well as rule-based methods, the first part of our project, which will be reported in our talk, will be to investigate the extent to what extent speech units are possible to single out for syntactic analysis, and to expand our rule-based phrase structure grammar to account for spoken sentence fragments. We are especially interested in looking into fragments that contains missing material in initial and medial position, and comparing these to those that have final material missing. We think it is likely that the former two types will be more systematic, since the missing material is a function of the speaker’s choice, as opposed to the latter, which can be
caused by speaker-external interruptions.

While we think Nygård’s analysis of sentence fragments is convincing, we do not think, as mentioned above, that it is feasible to follow that line in a computational analysis. Instead, we believe that the main sentence fragment types, including especially the exclamatory types mentioned by Teleman (1974), are finite in number. This line will be pursued, and sentence fragments that can possibly be analysed as sentences with missing material, will be regarded as trees in their own right:

(9) Neimen om hun visste!
    Not as if she knew

Other types of speech material will be regarded as part of larger sentence structures, either in three-dimensional structures (cf. Espinal 1991) or as adjunction.

References
The NoTa project: a corpus of spoken Norwegian.

http://www.hf.uio.no/tekstlab/prosjekter/NoTa/NoTa.htm