Annotating Information Structure in a Corpus of Spoken Danish

Patrizia Paggio

Centre for Language Technology
University of Copenhagen
patrizia@cst.dk

Abstract

This paper presents the work done to annotate a corpus of spoken Danish with information structure tags, and describes a preliminary study in which the corpus has been used to investigate the relation between focus and intra-clausal pauses. The study indicates that the pauses that do fall within the focus domain, tend to precede property-expressing words by which the object in focus is distinguished from other similar ones.

1. Introduction

This work has a two-fold motivation. Firstly, the importance and usefulness of enriching language corpora with information structure tags has recently been emphasised by several authors. For example, Kruijff-Korbayová and Kruijff (2004) propose a rich discourse-level annotation methodology to study information structure in corpora, while both Postolache (2005) and Diderichsen and Elming (2005) deal with the application of machine learning to the problem of automatic identification of topic and focus.

Secondly, an annotated corpus of spoken Danish which could be enriched with an information structure annotation layer has recently become available. This is the 'DanPASS' corpus, a collection of spoken monologues and dialogues that is described elsewhere in this volume (Grønnum, 2006). The corpus has been annotated with several annotation tiers, including orthographic and phonetic transcriptions, pauses, phonetic phrases and PoS-tags. Grønnum provides details on the corpus construction and the phonetic annotation tiers, as well as references to similar methods aiming at the elicitation of spontaneous speech. In this paper, we deal with the information structure annotation that is being added to a portion of the corpus, a collection of 54 monologues produced by 18 different subjects and dealing with three well-defined tasks. In the first task, the subjects describe a geometrical network, in the second they instruct a listener in assembling the drawing of a house out of existing pieces, and in the third they solve a map task. First we describe the methodology adopted to add an information structure tier to the corpus, and give an overview of the resulting resource. Then we briefly discuss[1] an example of how the annotation can be used to investigate the existence of systematic correspondences between the various annotation levels. The example in question deals with the relation between focusing and intra-clausal pauses.

2. Information Structure: Theory and Annotation Categories

In the theoretical framework that has guided the annotation of information structure, which is mainly inspired by Lambrecht (1994), a first basic distinction is made between sentence focus, which expresses non-presupposed information, and a presupposed background part. The focus is assumed to be obligatory, while the background is not. A second distinction is made between the sentence topic, defined as a referent or a set of referents about which the focus expresses pertinent information, and the comment, which consists of the focus and any other background information. A further assumption is that the topic can be singled out by means of the "*What about X*" test (Reinhart, 1981). Not all sentences, however, are construed as being about a topic, and in such cases the test will fail.

An example can be seen in the following short corpus excerpt dealing with the house assembling task, where the subject first tells the listener to pick up one of the square pieces available, and then explains what should be done with it. Once the square is established as an active referent in the discourse, it can be used as a sentence topic. In this and subsequent examples, focus expressions are rendered in small caps and topic expressions are in bold face. Words that are not marked in either way belong to the background.

(1) Så TAGER DU EN LILLE FIRKANT [...] Du lægger **den** MIDT PÅ TREKANTEN der fungerer som tag
'Then YOU TAKE A SMALL SQUARE [...] You put **it** IN THE MIDDLE OF THE TRIANGLE that functions as a roof.'

3. The Annotation Methodology

The purpose of the annotation work was to annotate all the words expressing the focus and, if a topic could be identified, those making up the topic. The background, on the contrary, was left uncoded. A set of written guidelines was developed based on the general principles mentioned in the preceding section, and two annotators were asked to tag the corpus. Two thirds of the corpus have been tagged by both annotators independently in order to evaluate and refine the guidelines. The last portion has been divided between the two coders, and is currently being annotated using the Praat tool (Boersma, 2001), which gives access to sound files and existing phonetic and orthographic transcriptions.

The annotation work relies largely on the coders' intuition, for example to decide what is presupposed information, whether a sentence referent can be unified with the 'X' in the topic test, or whether a subordinate clause like the relative *that functions as a roof* in example (1) should be included in the focus or not. The guidelines contain i. general definitions of focus and topic; ii. an explanation of the

[1] A more detailed exposition can be found in Paggio (2006).

topic test; and iii. a number of principles and heuristics for specific syntactic constructions or well-known ambiguous cases.

3.1. General Annotation Principles

General annotation principles define certain formal properties of topic and focus as they are used in this work, including their interrelation with syntactic and prosodic features. The principles are listed below.

- GENERAL
  Not all sentences have a topic.
  All sentences have a focus.
  Topic and focus are disjoint.

- SYNTACTIC PHRASES
  The focus need not be coincidental with a sentence phrase.

- FOCUS AND ACCENTUATION
  There must be at least one main accent in a focus domain, but there may be several.
  Words that are not part of the focus need not be deaccented, although they can be.

- DISCONTINUOUS FOCUS
  The focus domain may be discontinuous, i.e. contain a topic or other background, as in:

  (2) LÆG **den** DER.
  'PUT **it** THERE.'

The principles concerning accentuation merit some explanation. Unlike other languages, in Danish focus is not distinguished by a unique sentence accent. Danish prosody is characterised by the fact that content words like nouns and most verbs all carry an accent, regardless of whether they are in focus. Deaccentuation is mostly due to syntactic factors, although it can be used to convey a narrow focus or a contrast. In general, however, focus accent coincides with the rightmost accent in a sentence rather than the most prominent one, and non-presupposed information that is not in focus is not necessarily unaccented.

3.2. Specific Syntactic Environments

A number of principles concern syntactic environments where information structure is to a certain extent predictable. Below a few examples are shown.

- CLEFTS
  In a cleft, the focus is distributed between the cleft head and the tail:

  (3) Det er DEN der er LÆNGST
  lit: *It is* THAT *that is* LONGEST
  'That's the longest one'.

- EPISTEMIC CONSTRUCTIONS
  They often express the main content of the sentence in which they occur. In this (default) case, they contain topic and focus of the overall sentence, as in:

  (4) Det vil sige at **den** ER GUL.
  'Which means that **it** IS YELLOW.'

Other environments discussed in the guidelines are left dislocation, initial adverbials and copula verbs. In some cases the guidelines give a clear indication of how to treat the construction. For example, the resumptive pronoun in left dislocations is always coded as the sentence topic. In others, a clear-cut recipe does not exist, and several different examples from the corpus are given and discussed. An example are initial adverbials, which may be part of the focus or background. Intonation factors as well as givenness and discourse activation are all taken into consideration to decide in each specific case.

3.3. Annotation Heuristics

Heuristic principles are default rules that the annotators are invited to resort to when a more principled decision cannot be made. Examples are:

- INITIAL SENTENCE
  The initial sentence in a monologue is always interpreted as an all-focus construction. If the monologue consists of independent sections, there will be an initial sentence in each section.

- SUBORDINATE CLAUSES
  If a subordinate clause is treated as an independent sentence, focus and possibly topic are annotated. Otherwise, either it constitutes background information, in which case it is not annotated, or it is part of the focus domain in the matrix sentence, and then it is all annotated as focus.

Other heuristics concern the coding of fragments and repetitions. Finally, a preference for wide focus over narrow focus in cases of doubt has been added to the guidelines as a result of the intercoder agreement measurements discussed further below. This means that if an annotator cannot decide, e.g. based on intonation or discourse activation, whether the whole VP as opposed to only one of the complements belongs to the focus domain, the VP focus interpretation should be chosen.

4. The Annotated Corpus

The information structure annotation has been added to the DanPASS corpus in the form of an independent tier, where each of the words making up the topic or the focus is added a T or an F tag. An annotation example corresponding to a fragment of the example in (1) is shown below in the Praat textgrid format, where time stamps link the transcription to the corresponding sound file. Note that '+' corresponds to a pause, ' , ' is an accent, and '_' indicates the end of a word.

```
intervals [69]:
    xmin = 27.658522271545234
    xmax = 27.929271496109184
    text = "p_F"
intervals [70]:
    xmin = 27.929271496109184
    xmax = 28.19285472547374
    text = "+_"
```

```
intervals [71]:
    xmin = 28.19285472547374
    xmax = 28.890432172611025
    text = "tr,ekanten_F"
```

The same example is shown below in its entirety in a more compact linearised form, which we adopt from here on. Pauses are indicated by '+' and '='. The former is a silent pause, and the latter a pause accompanied by a sound, like 'hmm'.

(5) så t′ager du en/F l′ille/F f′irkant=/F du l′ægger den/T m′idt/F på/F + tr′ekanten/F der + fung′erer + som t′ag +
'Then YOU TAKE A SMALL SQUARE (*pause*) You put **it** IN THE MIDDLE OF (*pause*) THE TRIANGLE that (*pause*) functions (*pause*) as a roof (*pause*).'

	Focus	Topic	No tag	Total
Network C1	1608	268	2526	4402
Network C2	1889	287	2226	4402
House C1	4025	386	4151	8562
House C2	4193	377	3992	8562

Table 1: Tags in two corpus sections

So far, approximately two person months have been spent annotating the two sections of the corpus dealing with the network and the house tasks in two different versions. Table (1) shows the number of tags assigned by the two coders (C1 and C2). The kappa score varied between 0.7 to 0.8 depending on the corpus section, showing an acceptable inter-annotator agreement. Most disagreements relate to the identification of the focus left-hand boundary, where one of the annotators sometimes identified wider focus domains than the other, as in the following example:

(6) a. og v′induet/T h′ar/F = gr′ønne/F gard′iner/F i/F s′iden/F +
'And **the window** HAS GREEN CURTAINS ON THE SIDE.'

b. og v′induet/T h′ar = gr′ønne/F gard′iner/F i/F s′iden/F +
'And **the window** has GREEN CURTAINS ON THE SIDE.'

These discrepancies will be resolved by applying the guideline according to which, in cases of doubt, wide focus as in (a) should be preferred over narrow focus as in (b).

5. An Investigation: Focusing and Pauses

The annotated corpus, although not yet completed, gives us the possibility to investigate whether there are systematic relations between information structure and several prosodic and syntactic features. A preliminary study has already been carried out on the relation between focusing and pauses. The guidelines for the annotation of information structure, in fact, do not refer to pauses as a criterion for the annotators to take into consideration. Furthermore, the pauses indicated in the orthographic transcription have been recorded by a different set of annotators. Therefore, if a relation could be observed, it would not be due to prior bias.

Earlier studies (Jensen, 2005) (Hansen et al., 1993) have investigated the effect of syntactic boundaries (clausal as well as phrasal) on the placing of pauses in spoken Danish. In addition to the fact that there is a clear tendency for pauses to co-occur with clause boundaries, in both studies it is found that pauses falling within a syntactic phrase tend to occur towards the final part of the sentence. The authors make the hypothesis that this circumstance may be due to an effect related to focusing, since the final part of the sentence is also the locus of focused information. However, information structure is not annotated in the empirical material analysed in these works, and no clear conclusion can therefore be drawn on the issue.

The DanPASS corpus, on the other hand, gives us the means of testing this hypothesis. Thus a pilot study was carried out on a part of the resource (the network monologues, only in the version coded by one of the annotators) to verify i. whether intra-clausal pauses tend to be associated with the focus, and ii. where in the focus domain pauses tend to occur most frequently, for instance whether they mark the left-hand focus boundary. The reason why we only look at intra-clausal pauses is that we know already that most pauses mark clause boundaries.

The first question was investigated by comparing the frequency of occurrence of a pause before a focus word with the average frequency of a pause before any word. No distinction was made between silent and non-silent pauses. The number of words taken into account in the pilot study is 3659 words; the average pause frequency is 28.34%, and the frequency of a pause before a focus word is 20.29%, in other words significantly lower than the average. If we look at specific words within the focus domain, however, a more intersting pattern emerges. In the second part of the study, then, we investigated where in the focus domain pauses tend to occur.

Table (2) shows the frequency with which different part-of-speech categories occurring in the focus domain (i.e. tagged "F") are preceded by a pause. The total number of words considered is 1661. The figures show that pauses occur before adjectives significantly more often than before other word categories in the focus domain, and also more often than the average 28.34%. A slightly more precise characterisation of the occurrence of pauses in the focus domain was obtained by running a decision tree generator (Witten and Eibe, 2005) on the data. The two strongest rules learnt by the system (i.e. those with broadest coverage) predicted that i. a pause in the focus domain should be placed between a determiner and an adjective, and ii. a pause in the focus domain should be placed between an adjective and a noun. The two rules account for the two examples below.

(7) tilb′age er der en/F + r′ød/F f′irkant/F
'Left there is A (*pause*) RED SQUARE.'

(8) til v′enstre l′ægger du en/F r′ød/F + f′irkant/F
'To the left you put A RED (*pause*) SQUARE.'

	Adj	Adv	Conj	Det	N	Prep	Part	Pro	Verb	Other	Total
Pause	36.34	6.94	16.67	18.97	17.11	19.83	25.00	4.76	6.33	20.00	20.29
No pause	63.66	93.06	83.33	81.03	82.89	80.17	75.00	95.24	93.67	80.00	79.71
Total	100	100	100	100	100	100	100	100	100	100	100

Table 2: Distribution of pauses over part-of-speech categories in the focus domain (%)

Both examples – and the rules they give rise to – are quite characteristic of a recurrent communicative strategy in the monologues. In both network description and house construction tasks, in fact, the domain contains a number of geometrical figures which the various speakers have to tell apart either by means of their semantic type (a square rather than a triangle) or a distinctive property (colour, size, etc.). The pause in the focus domain, if there is a pause at all, tends to fall before the word that expresses this distinctive type or property. From the point of view of accentuation, however, this word is just as prominent as the other content words in focus, and is therefore not annotated as the only one in focus.

6. Discussion and Conclusion

In this paper we have described the methodology developed to add an annotation structure tier to the DanPASS corpus of spoken Danish. We have given examples of the guidelines used, shown annotation examples, and explained how the most frequent type of disagreement between the annotators has been resolved, resulting in an addition to the guidelines. The resulting resource opens up for very interesting investigation possibilities concerning the way in which information structure relates to prosodic and syntactic features.

In this paper the relation between pauses and focusing has been investigated by looking at the frequency with which pauses occur before words in focus compared to the average. It was found that taken individually, words in the focus domain have a lower probability of being preceded by a pause. By looking at words of specific syntactic categories, on the other hand, it appears that adjectives in the focus domain have a significantly higher probability of being preceded by a pause. This circumstance has been explained as a characteristic of the DanPASS corpus, according to which adjectives often serve the purpose of expressing the semantic property by which the domain object in focus is distinguished from other similar ones. In other words, at least in this corpus, pauses in the focus domain do not correspond to a phrasal boundary, nor do they mark the focus left-hand boundary. They express instead a more subtle relation with semantic features associated with the focused material.

It may be objected that looking at pauses before *words* is perhaps not the best way to capture the relation between pauses and focusing, if such a relation exists. A more significant picture may be revealed by treating the focus domain as a whole entity, and investigate whether there is a significant relation between pauses and focus domains rather than words. An interesting baseline for comparison could be the relation between pauses and prosodic phrases, since prosodic phrases according to Steedman (2003) correspond to information structural constituents. Since the DanPASS annotation also includes a tier for prosodic phrases, this investigation is an obvious possibility to pursue.

7. Acknowledgements

This work was supported by the Carlsberg Foundation.

8. References

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Philip Diderichsen and Jakob Elming. 2005. A corpus-based approach to topic in Danish dialog. In *Proceedings of the ACL Student Research Workshop*, pages 119–114. Ann Arbor Michigan, June.

Nina Grønnum. 2006. DanPASS - a Danish phonetically annotated spontaneous speech. In *Proceedings of LREC 2006*, Genova, Italy, May.

Peter Molbæk Hansen, Niels Reinholt Petersen, and Ebbe Spang-Hanssen. 1993. Syntactic boundaries and pauses in read-aloud Danish prose. In Björn Granström and Lennart Nord, editors, *Nordic Prosody VI. Papers from a symposium*, pages 159–172. Stockholm: Almqvist and Wiksell International.

Anne Jensen. 2005. *Clause Linkage in Spoken Danish*. Ph.D. thesis, Department of General and Applied Linguistics, University of Copenhagen, July.

Ivana Kruijff-Korbayová and Geert-Jan M. Kruijff. 2004. Discourse-level annotation for investigating information structure. In *Proceedings of the ACL Workshop on Discourse Annotation*, pages 41–48. Barcelona, Spain.

Knud Lambrecht. 1994. *Information Structure and Sentence Form*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Patrizia Paggio. 2006. Information structure and pauses in a corpus of spoken Danish. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics EACL*, Trento, Italy. In press.

Oana Postolache. 2005. Learning information structure in the Prague treebank. In *Proceedings of the ACL Student Research Workshop*, pages 115–120. Ann Arbor, Michigan, June.

Tanya Reinhart. 1981. Pragmatics and linguistics: an analysis of sentence topics. *Philosophica*, 27(1):53–94.

Mark Steedman. 2003. Information-structural semantics for English intonation. In *Proceedings of LSA Summer Institute Workshop on Topic and Focus*. Santa Barbara, July 2001.

Ian H. Witten and Frank Eibe. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann: San Francisco, 2nd edition.