Information structure and pauses in a corpus of spoken Danish

Patrizia Paggio
Centre for Language Technology
University of Copehagen
Denmark
patrizia@cst.dk

Abstract

This paper describes a study in which a corpus of spoken Danish annotated with focus and topic tags was used to investigate the relation between information structure and pauses. The results show that intra-clausal pauses in the focus domain, tend to precede those words that express the property or semantic type whereby the object in focus is distinguished from other ones in the domain.

1 Introduction

The interest for corpora annotated with information structure has been raised recently by several authors. Kruijff-Korbayová and Kruijff (2004) describe a method where a rich discourse-level annotation is used to investigate information structure, while both Postolache (2005) and Diderichsen and Elming (2005) study the application of machine learning to the problem of automatic identification of topic and focus. In this study, on the contrary, information structure is annotated manually, and the annotation is used to investigate the correlation between information structure tags and intra-clausal pauses.

2 Annotating information structure

The starting point for this study was the corpus of spoken Danish 'DanPass' (Grønnum, 2005), a collection of 54 monologues produced by 18 different subjects dealing with three well-defined tasks, following the methodology established in Terken (1985). In the first task, the subjects describe a geometrical network, in the second the process of assembling the drawing of a house out of existing pieces, and in the third they solve a map task. The corpus has been annotated with several annotation tiers, including orthography, phonetic transcription, pauses and PoS-tags. Two independent annotators added then tags for focus and topic based on a set of simple guidelines, and using the Praat tool to carry out the annotation.

The annotation reflects the assumption that a sentence can be divided into an obligatory focus part, which expresses the non-presupposed information, and a presupposed background part. A referent in the background part may function as the sentence topic in the sense of Lambrecht (1994). For each sentence in the corpus, the annotators were asked to identify what they intuitively considered non-presupposed information and annotate it as belonging to the focus. Technically, each word belonging to the focus is added a focus tag. The annotators were also asked to test whether they could single out a sentence referent by means of the "What about X" test (Reinhart, 1981). If they could, they were asked to add topic tags to all the words making up the corresponding expression. Words not bearing any tag are considered part of the background.

The guidelines did not contain any reference to pausing, nor did the annotators know that their work would be used to study the correlation be-

	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

tween pauses and information structure. In fact, that was not the purpose of the annotation work, which is of more general interest. It should also be noted that the annotators were not explicitly instructed to code *phrases*, since we did not want to make the assumption that topic or focus necessarily correspond to syntactic phrases. Approximately two person months were spent annotating two sections of the corpus. 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. These differences have not been inspected yet, but will be used to revise the guidelines to produce a unique consistent annotation. Table (1) shows the number of tags assigned by the two coders (C1 and C2) in the two sections of the corpus coded so far.

Below, an example of an annotated tier is shown in a linearised format (the textgrids output by Praat also contain time intervals that link the transcription to the sound file):

(1) + ovenover + er der en/F + grøn/F cirkel/F = og oven over den/T grønne/T cirkel/T er der en/F + lilla/F trekant/F +
'PAUSE above PAUSE there is [$_F$ a PAUSE green circle] PAUSE and above [$_T$ the green circle] there is [$_F$ a PAUSE purple triangle]'

The example consists of two sentences. In the first, the annotator has tagged 'en grøn cirkel' (a green circle) as the focus; in the second, 'den grønne cirkel' (the green circle) has been tagged as the topic, while 'en lilla trekant' (a purple triangle) is tagged as the focus. Pauses are indicated by '+' and '='. The former is a silent pause, and the latter a pause accompanied by a sound, like 'hmm'. Pauses were already available in the orthographic transcription of the corpus, which was produced earlier by different annotators.

3 Pauses in earlier studies

The material annotated so far already gives us the possibility to investigate whether there is a significant relation between pauses and information structure. Earlier studies (Jensen, 2005) (Hansen et al., 1993) investigated the effect of syntactic boundaries (clausal as well as phrasal) on the placing of pauses in spoken Danish. In the first study, it is found that more than 55% of the pauses co-occur with clause boundaries, 12% with phrase boundaries, and the remaining 33% occur within phrases or in conjunction with repairs, interjections and enumerations. It is also noted that pauses falling within a syntactic phrase tend to be placed in the final part of the sentence. The second study confirms this observation by showing that 60% of the pauses that do not co-occur with syntactic boundaries occur within the last 40% of the sentence (measured in number of syllables). The authors of both investigations make the hypothesis that information structure may have an effect on the occurrence of pauses within clauses. However, the empirical material used in those works is not annotated with respect to information structure, and therefore, no conclusive claim could be made. In addition, the data used in Hansen *et al* (1993) come from news reading, and are thus essentially written language although delivered orally.

4 Pauses and focusing in Danish

The purpose of this pilot study is, on the basis of the annotated DanPass corpus, to verify i. to what degree pauses tend to be associated with focus and topic, and ii. where in the focus domain pauses tend to occur, particularly whether pauses are used to mark the left-hand focus boundary.

Since we already know from the studies cited above that there is a strong tendency for pauses to coincide with clause boundaries, we decided

	F word	T word	No tag	Total
Pause	20.29	7.59	39.70	28.34
No pause	79.71	92.41	60.30	71.66
Total	100	100	100	100

Table 2: Distribution of pauses over information structure categories (%)

to exclude those from the study, and only look at pauses that occur within clauses. So far, the investigation has been carried out for the network description part of the corpus, and only for the data produced by one of the coders.

The first question – whether pauses relate to words coded as either focus or topic – was investigated by counting, out of a total 3659 words, how many words tagged as either F or T, or bearing no tag, are preceded by a pause (silent or non silent). The results, shown in Tables (2), seem to disconfirm the hypothesis that there should be a correlation between pauses and information structure categories, or at least that a correlation, if it exists, can be expressed by looking at the frequency with which pauses precede focus or topic *words*. In fact, over 65% of the intra-clausal pauses in the material precede untagged words, and the observed frequency of a pause before a focus or a topic word is lower than the average 28.34% (baseline).

Since we know that topics often occur sentence-initially, the results in the tables are misleading in that *only* intra-clausal pauses are taken into consideration. Therefore we also looked at what percentage of topic words are succeeded rather than preceded by a pause, and found that 33.50% are. This figure is interesting, but needs further investigations.

Now we zoom in on the focus domain. First of all, we look at pause distribution across different part-of-speech categories, again by inspecting the pauses *preceding* words. Table (3) 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 no. of words considered is 1661.

The interesting fact that emerges is that adjectives have a remarkably higher probability to be preceded by a pause than any of the other category, and also a clearly higher probability than the average 28.34%.

We then looked at the first pause in the focus domain. The first pause falls before the first focus word in only 30% of the cases. In other words, it does not seem to mark the left-hand boundary of the focus domain. By running a decision tree generator (Witten and Eibe, 2005) on the data, we found that the strongest rule learnt by the system was one that places the first pause in the focus domain between a determiner and an adjective (2). Another rule predicts that a pause will fall between an adjective and a noun (3).

(2) tilbage er der... en/F + rød/F firkant/F
'left there is... [$_F$ a PAUSE red square]'

(3) til venstre... lægger du en/F rød/F + firkant/F
'to the left... you put [$_F$ a red PAUSE square]'

The two rules reflect a strong characteristic of the monologues under investigation, where the speakers have to draw the listener's attention to the various geometrical figures in the network they are describing. To tell them apart from each other, they either use the colour of the figure or its shape. In other words, the pauses occurring in the focus domain tend to precede the word that expresses what Dik (1989) calls selecting focus, here an adjective that, by defining a selecting property or type, helps distinguishing the object in focus from other similar ones. From the point of view of accentuation, however, the adjective is not more prominent than the noun, and is therefore not annotated as the only word in focus.

5 Conclusions and further research

In conclusion, the pilot study shows that words making up the topic or the focus of a sentence do not show a general tendency to be preceded by pauses. However, preliminary results indicate that topics tend to be followed by pauses. Furthermore, words belonging to specific syntatic

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

categories may have a significantly higher probability to be preceded by a pause than a randomly chosen word. In the corpus we have worked with, these words express the property or semantic type whereby the object in focus can be distinguished from other similar objects. In other words, the system by which Danish speakers use pauses seems sensitive to information structure in a subtle way that, at least as far as focus is concerned, creates boundaries that do not necessarily correspond to those between syntactic constituents.

An interesting issue we haven't yet addressed is whether intra-clausal pauses relate to prosodic phrases, which according to Steedman (2001) correspond to information structural constituents. Since the DanPass annotation also foresees a tier for prosodic phrases, this investigation is possible. Furthermore, we want to test whether there are differences in the way in which different users relate pauses to topic establishment and focusing. We know already now that the percentage of pauses per word varies across speakers, and that speakers' individual pause rates do not vary much depending on the task. The corpus provides a very nice means of studying whether they use pauses for different purposes.

Acknowledgements

This work was supported by the Carlsberg Foundation.

References

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.

Simon Dik. 1989. *The Theory of Functional Grammar*. Functional Grammar Series. Dordrecht: Foris Publications.

Nina Grønnum. 2005. DanPASS - Danish phonetically annotated spontaneous speech. Talk given at FONETIK 2005 in Gothemburg, 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*.

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

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

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

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

Jacques M. B. Terken. 1985. *Use and Function of Accentuation: Some Experiments*. Ph.D. thesis, Leiden University, September.

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