A corpus-based approach to topic in Danish dialog*

Philip Diderichsen
Lund University Cognitive Science
Lund University
Sweden
philip.diderichsen@lucs.lu.se

Jakob Elming
CMOL / Dept. of Computational Linguistics
Copenhagen Business School
Denmark
je.id@cbs.dk

Abstract

We report on an investigation of the pragmatic category of topic in Danish dialog and its correlation to surface features of NPs. Using a corpus of 444 utterances, we trained a decision tree system on 16 features. The system achieved near-human performance with success rates of 84–89% and F_1-scores of 0.63–0.72 in 10-fold cross validation tests (human performance: 89% and 0.78). The most important features turned out to be preverbal position, definiteness, pronominalisation, and non-subordination. We discovered that NPs in epistemic matrix clauses (e.g. "I think . . . ") were seldom topics and we suspect that this holds for other interpersonal matrix clauses as well.

1 Introduction

The pragmatic category of topic is notoriously difficult to pin down, and it has been defined in many ways (Büring, 1999; Davison, 1984; Engdahl and Vallduví, 1996; Gundel, 1988; Lambrecht, 1994; Reinhart, 1982; Vallduví, 1992). The common denominator is the notion of topic as what an utterance is about. We take this as our point of departure in this corpus-based investigation of the correlations between linguistic surface features and pragmatic topicality in Danish dialog.

Danish is a verb-second language. Its word order is fixed, but only to a certain degree, in that it allows any main clause constituent to occur in the preverbal position. The first position thus has a privileged status in Danish, often associated with topicality (Harder and Poulsen, 2000; Togeby, 2003). We were thus interested in investigating how well the topic correlates with the preverbal position, along with other features, if any.

Our findings could prove useful for the further investigation of local dialog coherence in Danish. In particular, it may be worthwile in future work to study the relation of our notion of topic to the C_b of Grosz et al.s (1995) Centering Theory.

2 The corpus

The basis of our investigation was two dialogs from a corpus of doctor-patient conversations (Hermann, 1997). Each of the selected dialogs was between a woman in her thirties and her doctor. The doctor was the same in the two conversations, and the overall topic of both was the weight problems of the patient. One of the dialogs consisted of 125 utterances (165 NPs), the other 319 (449 NPs).

3 Method

The investigation proceeded in three stages: first, the topic expressions (see below) of all utterances were identified[1]; second, all NPs were annotated for linguistic surface features; and third, decision trees

*We thank Daniel Hardt and two anonymous reviewers for many helpful comments on drafts of this paper.

[1] Utterances with dicourse regulating purpose (e.g. yes/no-answers), incomplete utterances, and utterances without an NP were excluded.

Proceedings of the ACL Student Research Workshop, pages 109–114,
Ann Arbor, Michigan, June 2005.

were generated in order to reveal correlations between the topic expressions and the surface features.

3.1 Identification of topic expressions

Topics are distinguished from *topic expressions* following Lambrecht (1994). Topics are entities pragmatically construed as being what an utterance is about. A topic expression, on the other hand, is an NP that formally expresses the topic in the utterance. Topic expressions were identified through a two-step procedure; 1) identifying topics and 2) determining the topic expressions on the basis of the topics.

First, the topic was identified strictly based on pragmatic aboutness using a modified version of the '*about* test' (Lambrecht, 1994; Reinhart, 1982).

The *about* test consists of embedding the utterance in question in an 'about-sentence' as in Lambrecht's example shown below as (1):

(1) He said about the children that they went to school.

This is a paraphrase of the sentence *the children went to school* which indicates that the referent of *the children* is the topic because it is appropriate (in the imagined discourse context) to embed this referent as an NP in the *about* matrix clause. (Again, the referent of *the children* is the topic, while the NP *the children* is the topic expression.)

We adapted the *about* test for dialog by adding a request to 'say something about ...' or 'ask about ...' before the utterance in question. Each utterance was judged in context, and the best topic was identified as illustrated below. In example (2), the last utterance, (2-D_3), was assigned the topic TIME OF LAST WEIGHING. This happened after considering which *about* construction gave the most coherent and natural sounding result combined with the utterance. Example (3) shows a few *about* constructions that the coder might come up with, and in this context (3-iv) was chosen as the best alternative.

(2) D_1 sid ned og lad mig høre, Annette (made-up name)
sit down and let me hear, Annette
P_1 jeg skal bare vejes
I shall just be.weighed
P_2 og så skal jeg have svar fra sidste gang
and then shall I have answer from last time
D_2 så skal vi se en gang
then let us see one time
D_3 **det**... er... fjorten dage siden du blev vejet...
it... is... fourteen days since you were weighed...

(3) i. Say something about THE PATIENT (=you).
ii. Say something about THE WEIGHING OF THE PATIENT.
iii. Say something about THE LAST WEIGHING OF THE PATIENT.
iv. Say something about THE TIME OF LAST WEIGHING OF THE PATIENT.

Creating the *about* constructions involved a great deal of creativity and made them difficult to compare. Sometimes the coders chose the exact same topic, at other times they were obviously different, but frequently it was difficult to decide. For instance, for one utterance Coder 1 chose OTHER CAUSES OF EDEMA SYMPTOM, while Coder 2 chose THE EDEMA'S CONNECTION TO OTHER THINGS. Slightly different wordings like these made it impossible to test the intersubjectivity of the topic coding.

The second step consisted in actually identifying the topic expression. This was done by selecting the NP in the utterance that was the best formal representation of the topic, using 3 criteria:

1. The topic expression is the NP in the utterance that refers to the topic.
2. If no such NP exists, then the topic expression is the NP whose referent the topic is a property or aspect of.
3. If no NP fulfills one of these criteria, then the utterance has no topic expression.

In the example from before, (2-D_3), it was judged that *det* 'it' (emphasized) was the topic expression of the utterance, because it shared reference with the chosen topic from (3-iv).

If two NPs in an utterance had the same reference, the best topic representative was chosen. In reflexive constructions like (4), the non-reflexive NP, in this case *jeg* 'I', is considered the best representative.

(4) men jeg har ikke tabt mig
but I have not lost me (i.e. lost weight)

In syntactially complex utterances, the best representative of the topic was considered the one occurring in the clause most closely related to the topic. In the following example, since the topic was THE PATIENT'S HANDLING OF EATING, the topic expression had to be one of the two instances of *jeg* 'I'. Since the topic arguably concerns 'handling' more than 'eating', the NP in the matrix clause (emphasized) is the topic expression.

(5) **jeg** har slet ikke tænkt på hvad jeg har spist
I have really not thought about what I have eaten

A final example of several NPs referring to the same topic has to do with left-dislocation. In example (6), the preverbal object *ham* 'him' is immediately preceded by its antecedent *min far* 'my father'. Both NPs express the topic of the utterance. In Danish, resumptive pronouns in left-dislocation constructions always occur in preverbal position, and in cases where they express the topic there will thus always be two NPs directly adjacent to each other which both refer to the topic. In such cases, we consider the resumptive pronoun the topic expression, partly because it may be considered a more integrated part of the sentence (cf. Lambrecht (1994)).

(6) min far **ham** så jeg sjældent
my father him saw I seldom

The intersubjectivity of the topic expression annotation was tested in two ways. First, all the topic expression annotations of the two coders were compared. This showed that topic expressions can be annotated reasonably reliably (κ = 0.70 (see table 1)). Second, to make sure that this intersubjectivity was not just a product of mutual influence between the two authors, a third, independent coder annotated a small, random sample of the data for topic expressions (50 NPs). Comparing this to the annotation of the two main coders confirmed reasonable reliability (κ = 0.70).

3.2 Surface features

After annotating the topics and topic expressions, 16 grammatical, morphological, and prosodic features were annotated. First the smaller corpus was annotated by the two main coders in collaboration in order to establish annotating policies in unclear cases. Then the features were annotated individually by the two coders in the larger corpus.

Grammatical roles. Each NP was categorized as grammatical subject (sbj), object (obj), or oblique (obl).These features can be annotated reliably (sbj: C1 (number of sbj's identified by Coder 1) = 208, C2 (sbj's identified by Coder 2) = 207, C1+2 (Coder 1 and 2 overlap) = 207, κ_{sbj} = **1.00**; obj: C1 = 110, C2 = 109, C1+2 = 106, κ_{obj} = **0.97**; obl: C1 = 30, C2 = 50, C1+2 = 29, κ_{obl} = **0.83**).

Morphological and phonological features. NPs were annotated for pronominalisation (pro), definiteness (def), and main stress (str). (Note that the main stress distinction only applies to pronouns in Danish.) These can also be annotated reliably (pro: C1 = 289, C2 = 289, C1+2 = 289, κ_{pro} = **1.00**; def: C1 = 319, C2 = 318, C1+2 = 318, κ_{def} = **0.99**; str: C1 = 226, C2 = 226, C1+2 = 203, κ_{str} = **0.80**).

Unmarked surface position. NPs were annotated for occurrence in pre-verbal (pre) or post-verbal (post) position relative to their subcategorizing verb. Thus, in the following example, *det* 'it' is +pre, but –post, because *det* is not subcategorized by *tror* 'think'.

(7) Ø tror [+pre,–post **det**] hjælper lidt
(I) think [+pre,–post it] helps a little

In addition to this, NPs occurring in pre-verbal position were annotated for whether they were repetitions of a left-dislocated element (ldis). Example (8) further exemplifies the three position-related features.

(8) min far [+ldis,+pre ham] så [+post jeg] sjældent
my father [+ldis,+pre him] saw [+post I] seldom

All three features can be annotated highly reliably (pre: C1 = 142, C2 = 142, C1+2 = 142, κ_{pre} = **1.00**; post: C1 = 88, C2 = 88, C1+2 = 88, κ_{post} = **1.00**; ldis: C1 = 2, C2 = 2, C1+2 = 2, κ_{ldis} = **1.00**).

Marked NP-fronting. This group contains NPs fronted in marked constructions such as the passive (pas), clefts (cle), Danish 'sentence intertwining' (dsi), and XVS-constructions (xvs).

NPs fronted as subjects of passive utterances were annotated as +pas.

(9) [+pas **jeg**] skal bare vejes
[+pas I] shall just be.weighed

A cleft construction is defined as a complex construction consisting of a copula matrix clause with a relative clause headed by the object of the matrix clause. The object of the matrix clause is also an argument or adjunct of the relative clause predicate. The clefted element *det* 'that', which we annotate as +cle, leaves an 'empty slot', *e*, in the relative clause, as shown in example (10):

(10) det er jo ikke [+cle **det**$_i$] du skal tabe dig
it is after all not [+cle that$_i$] you shall lose weight
af ***e***$_i$ som sådan
from e_i as such

Danish sentence intertwining can be defined as a special case of extraction where a non-WH constituent of a subordinate clause occurs in the first

position of the matrix clause. As in cleft constructions, an 'empty slot' is left behind in the subordinate clause. NPs in the fronted position were annotated as +dsi:

(11) [+dsi **det**$_i$] tror jeg ikke det gør $\boldsymbol{e}_i$
[+dsi that$_i$] think I not it does e_i

The XVS construction is defined as a simple declarative sentence with anything but the subject in the preverbal position. Since only one constituent is allowed preverbally[2], the subject occurs after the finite verb. In example (12), the finite verb is an auxiliary, and the canonical position of the object after the main verb is indicated with the 'empty slot' marker *e*. The preverbal element in XVS-constructions is annotated as +xvs.

(12) [+xvs **det**$_i$] har jeg altså haft $\boldsymbol{e}_i$ før
[+xvs that$_i$] have I truly had e_i before

All four features can be annotated highly reliably (pas: C1 = 1, C2 = 1, C1+2 = 1, κ_{pas} = **1.00**; cle: C1 = 4, C2 = 4, C1+2 = 4, κ_{cle} = **1.00**; dsi C1 = 3, C2 = 3, C1+2 = 3, κ_{dsi} = **1.00**; xvs: C1 = 18, C2 = 18, C1+2 = 18, κ_{xvs} = **1.00**).

Sentence type and subordination. Each NP was annotated with respect to whether or not it appeared in an interrogative sentence (int) or a subordinate clause (sub), and finally, all NPs were coded as to whether they occurred in an epistemic matrix clause or in a clause subordinated to an epistemic matrix clause (epi). An epistemic matrix clause is defined as a matrix clause whose function it is to evaluate the truth of its subordinate clause (such as "*I think* ..."). The following example illustrates how we annotated both NPs in the epistemic matrix clause and NPs in its immediate subordinate clause as +epi, but not NPs in further subordinated clauses. The +epi feature requires a +/–sub feature in order to determine whether the NP in question is in the epistemic matrix clause or subordinated under it. Subordination is shown here using parentheses.

(13) [+epi jeg] tror mere ([+epi,+sub det] er fordi (at
[+epi I] think rather ([+epi,+sub it] is because (that
[+sub man] spiser på [+sub dumme tidspunkter] ik'))
[+sub you] eat at [+sub stupid times] right))

All features in this group can be annotated reliably (int: C1 = 55, C2 = 55, C1+2 = 55, κ_{int} = **1.00**; sub: C1 = 117, C2 = 111, C1+2 = 107, κ_{sub} = **0.93**; epi: C1 = 38, C2 = 45, C1+2 = 37, κ_{epi} = **0.92**).

3.3 Decision trees

In the third stage of our investigation, a decision tree (DT) generator was used to extract correlations between topic expressions and surface features. Three different data sets were used to train and test the DTs, all based on the larger dialog.

Two of these data sets were derived from the complete set of NPs annotated by each main coder individually. These two data sets will be referred to below as the 'Coder 1' and 'Coder 2' data sets.

The third data set was obtained by including only NPs annotated identically by both main coders in relevant features[3]. This data set represents a higher degree of intersubjectivity, especially in the topic expression category, but at the cost of a smaller number of NPs. 63 out of a total of 449 NPs had to be excluded because of inter-coder disagreement, 50 due to disagreement on the topic expression category. This data set will be referred to below as the 'Intersection' data set.

A DT was generated for each of these three data sets, and each DT was tested using 10-fold cross validation, yielding the success rates reported below.

4 Results

Our results were on the one hand a subset of the features examined that correlated with topic expressions, and on the other the discovery of the importance of different types of subordination. These results are presented in turn.

4.1 Topic-indicating features

The optimal classification of topic expressions included a subset of important features which appeared in every DT, i.e. +pro, +def, +pre, and –sub. Several other features occurred in some of the DTs, i.e. dsi, int, and epi. The performance of all the DTs is summarized in table 2 below.

[2] Only one constituent is allowed in the *intrasentential* preverbal position. Left-dislocated elements are not considered part of the sentence proper, and thus do not count as preverbal elements, cf. Lambrecht (1994).

[3] "Relevant features" were determined in the following way: A DT was generated using a data set consisting only of NPs annotated identically by the two coders in all the features, i.e. the 16 surface features as well as the topic expression feature. The features constituting this DT, i.e. pro, def, sub, and pre, as well as the topic expression category, were relevant features for the third data set, which consisted only of NPs coded identically by the two coders in these 5 features.

The DT for the Coder 1 data set contains the features def, pro, dsi, sub, and pre. According to this classification, a definite pronoun in the fronted position of a Danish sentence intertwining construction is a topic expression, and other than that, definite pronouns in the preverbal position of non-subordinate clauses are topic expressions. The 10-fold cross validation test yields an 84% success rate. F_1-score: 0.63.

The Coder 2 DT contains the features pro, def, sub, pre, int, and epi. Here, if a definite pronoun occurs in a subordinate clause it is not a topic expression, and otherwise it is a topic expression if it occurs in the preverbal position. If it does not occur in preverbal position, but in a question, it is also a topic expression unless it occurs in an epistemic matrix clause. Success rate: 85%. F_1-score: 0.67.

Finally, the Intersection DT contains the features pro, def, sub, and pre. According to this DT, only definite pronouns in preverbal position in non-subordinate clauses are topic expressions. The DT has a high success rate of 89% in the cross validation test — which is not surprising, given that a large number of possibly difficult cases have been removed (mainly the 50 NPs where the two coders disagreed on the annotation of topic expressions). F_1-score: 0.72.

Since there is no gold standard for annotating topic expressions, the best evaluation of the human performance is in terms of the amount of agreement between the two coders. Success rate and F_1 analogs for human performance were therefore computed as follows, using the figures displayed in table 1.

		Coder 2		Total
		Topic	Non-topic	
Coder 1	Topic	88	27	115
	Non-topic	23	311	334
Total		111	338	449

Table 1: The topic annotation of Coder 1 and Coder 2.

Success rate analog: The agreement percentage between the human coders when annotating topic expressions ($\frac{449 \text{ NPs} - (23+27) \text{ NPs}}{449 \text{ NPs}} \times 100 = 89\%$).

***F*$_1$ analog**: The performance of Coder 1 evaluated against the performance of Coder 2 ("Precision": $\frac{88}{88+27} = 0.77$; "Recall": $\frac{88}{88+23} = 0.79$; "F_1": $2 \times \frac{0.77 \times 0.79}{0.77+0.79} = 0.78$).

Data set	Coder 1	Coder 2	Intersect.	Human
Total NPs	449	449	386	449
Success rate	84%	85%	89%	89%
Precision	0.77	0.74	0.79	0.79
Recall	0.53	0.61	0.67	0.77
F_1-score	0.63	0.67	0.72	0.78

Table 2: Success rates, Precision, Recall, and F_1-scores for the three different data sets. For comparison, we added success rate and F_1 analogs for human performance.

4.2 Interpersonal subordination

We found that syntactic subordination does not have an invariant function as far as information structure is concerned. The emphasized NPs in the following examples are definite pronouns in preverbal position in syntactically non-subordinate clauses. But none of them are perceived as topic expressions.

(14) så **det** kan godt være at hvis man har... tabt noget
so it may well be that if you have... lost some
mere i løbet af ugen ik'
more during the.week right

(15) **jeg** tror mere det er fordi at man spiser på
I think rather it is because that you eat at
dumme tidspunkter ik'
stupid times right

The reason seems to be that these NPs occur in epistemic matrix clauses (+epi).

The following utterances have not been annotated for the +epi feature, since the matrix clauses do not seem to state the speaker's attitude towards the truth of the subordinate clause. However, the emphasized NPs seem to stand in a very similar relation to the message being conveyed, and none of them were perceived as topic expressions.

(16) men altså **jeg** har bare bemærket at at det
but you know I have just noticed that that it
er blevet værre ik'
has become worse right

(17) og **det** kan man da sige på tre uger det er
and that can you though say in three weeks that is
da ikke vildt meget
surely not wildly much

This suggests that a more general type of matrix clause than the epistemic matrix clause, namely the *interpersonal matrix clause* (Jensen, 2003) would be relevant in this context. This category would cover all of the above cases. It is defined as a matrix clause that expresses some attitude towards the mes-

sage conveyed in its subordinate clause. This more general category presumably signals non-topicality rather than topicality just like the special case of epistemic subordination.

5 Summary and future work

We have shown that it is possible to generate algorithms for Danish dialog that are able to predict the topic expressions of utterances with near-human performance (success rates of 84–89%, F_1 scores of 0.63–0.72).

Furthermore, our investigation has shown that the most characteristic features of topic expressions are preverbal position (+pre), definiteness (+def), pronominal realisation (+pro), and non-subordination (–sub). This supports the traditional view of topic as the constituent in preverbal position.

Most interesting is subordination in connection with certain matrix clauses. We discovered that NPs in epistemic matrix clauses were seldom topics. In complex constructions like these the topic expression occurs in the subordinate clause, not the matrix clause as would be expected. We suspect that this can be extended to the more general category of inter-personal matrix clauses.

Future work on dialog coherence in Danish, particularly pronoun resolution, may benefit from our results. The centering model, originally formulated by Grosz et al. (1995), models discourse coherence in terms of a 'local center of attention', viz. the *backward-looking center*, C_b. Insofar as the C_b corresponds to a notion like topic, the corpus-based investigation reported here might serve as the empirical basis for an adaptation for Danish dialog of the centering model. Attempts have already been made to adapt centering to dialog (Byron and Stent, 1998), and, importantly, work has also been done on adapting the centering model to other, freer word order languages such as German (Strube and Hahn, 1999).

References

Daniel Büring. 1999. Topic. In Peter Bosch and Rob van der Sandt, editors, *Focus — Linguistic, Cognitive, and Computational Perspectives*, pages 142–165. Cambridge University Press.

Donna K. Byron and Amanda J. Stent. 1998. A preliminary model of centering in dialog. Technical report, The University of Rochester.

Alice Davison. 1984. Syntactic markedness and the definition of sentence topic. *Language*, 60(4).

Elisabeth Engdahl and Enric Vallduví. 1996. Information packaging in HPSG. *Edinburgh working papers in cognitive science: Studies in HPSG*, 12:1–31.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Jeanette K. Gundel. 1988. Universals of topic-comment structure. In Michael Hammond, Edith Moravcsik, and Jessica Wirth, editors, *Studies in syntactic typology*, volume 17 of *Studies in syntactic typology*, pages 209–239. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Peter Harder and Signe Poulsen. 2000. Editing for speaking: first position, foregrounding and object fronting in Danish and English. In Elisabeth Engberg-Pedersen and Peter Harder, editors, *Ikonicitet og struktur*, pages 1–22. Netværk for funktionel lingvistik, Copenhagen.

Jesper Hermann. 1997. Dialogiske forståelser og deres grundlag. In Peter Widell and Mette Kunøe, editors, *6. møde om udforskningen af dansk sprog*, pages 117–129. MUDS, Århus.

K. Anne Jensen. 2003. *Clause Linkage in Spoken Danish*. Ph.D. thesis from the University of Copenhagen, Copenhagen.

Knud Lambrecht. 1994. *Information structure and sentence form: topic, focus and the mental representations of discourse referents*. Cambridge University Press, Cambridge.

Tanya Reinhart. 1982. Pragmatics and linguistics. an analysis of sentence topics. *Distributed by the Indiana University Linguistics Club.*, pages 1–38.

Michael Strube and Udo Hahn. 1999. Functional centering — grounding referential coherence in information structure. *Computational linguistics*, 25(3):309–344.

Ole Togeby. 2003. *Fungerer denne sætning? – Funktionel dansk sproglære*. Gads forlag, Copenhagen.

Enric Vallduví. 1992. *The informational component*. Ph.D. thesis from the University of Pennsylvania, Philadelphia.