Semi-automatic Identification of Danish Discourse Deictics

Costanza Navarretta

Center for Language Technology
Njalsgade 80, 2300 Copenhagen S - DK
tel. 45 + 35 32 90 65
fax. 45 + 35 32 90 89
costanza@cst.ku.dk

Abstract. In this paper we present an algorithm for the (semi-)automatic identification of anaphors whose antecedents are verbal phrases, clauses or discourse segments in Danish Dialogues. Although these anaphors are quite frequent, especially in conversations, they are usually been neglected in computational linguistics. The algorithm we propose contains defeasible rules for distinguishing these anaphors from those who have individual nominals as antecedents. The rules have been identified by looking at the occurrences of these types of anaphor in the transcriptions of two dialogue collections. The algorithm has been manually tested on four Danish dialogues and the obtained results have been evaluated.[1]

1 Introduction

This paper deals with the identification of Danish anaphors whose antecedents are verbal phrases, clauses and discourse segments. Following [18], we call them discourse deictics.[2] Two examples of Danish discourse deictics are given in (1).

(1) *og så prøvede jeg så at gå lidt i svømmehallen (1 sek) og* **det** *prøver jeg sådan ind imellem, men jeg hader* **det**
(lit. and then tried I then to go a little to the swimming pool (1 sec.) and **it** try I such from time to time, but I hate **it**)
(and then I tried to go a little to the swimming pool (1 sec.) and I still try from time to time, but I hate **it**)

The two occurrences of the pronoun *det* (it/this/that) in (1) refer to the infinitive *at gå i svømmehallen* (to go to the swimming pool).

Although discourse deictics occur very frequently, especially in spoken language, they are seldom dealt with in literature, because their treatment is quite problematic. First of all it is difficult to distinguish them from pronouns with non-abstract antecedents because the same pronouns are used to refer to both abstract and non-abstract entities. It is also hard to recognise the correct antecedent, i.e. verbal phrases, clauses or discourse

[1] The research described has been done under the Staging project which is funded by the Danish Research Councils.

[2] Anaphors with nominal antecedents having an abstract referent can be included in the group of discourse deictics, but we have not looked at them in the present work.

V. Matoušek et al. (Eds.): TSD 2001, LNAI 2166, pp. 396–402, 2001.
© Springer-Verlag Berlin Heidelberg 2001

segments. Finally the semantic object pointed to by the deictic must be identified, see i.a. [13,18]. Despite these difficulties it is important to identify discourse deictics because they cannot be treated as anaphors referring to non-abstract objects. This paper deals with this aspect.

We have based our study of discourse deictics on their occurrences in the transcriptions of two Danish dialogue collections, *Samtale hos Lægen* ("The Talk at the Doctor's"), henceforth **SL**, [4,12] and **BySoc** [9,11]. The conversations have been collected by researchers at the University of Copenhagen and contain approx. 89,000 and one million running words, respectively. We have supplied our research by looking at the occurrences of discourse deictics in the written text corpus, **Bergenholtz** [2] containing approx. five million words.

In section 2 we present the Danish data. In section 3 we discuss the background for our work and we propose preference rules for identifying Danish discourse deictics. In section 4 we evaluate these rules while in section 5 we make some concluding remarks.

2 Danish Discourse Deictics

Discourse deictics in Danish comprise the following third-person neuter gender pronouns and demonstratives: *det* (it, this, that), *dette* (this), *det her* (this) and *det der* (that). The most common discourse deictic is *det*, while *dette* is mostly used in written language. We only found one occurrence of it in our dialogue collections.

Examples of Danish discourse deictics are the following:

- discourse deictic corefers with a clause:

  (2) **A:** *Du skal tage en blodprøve*
  (You have to take a blood test)
  B: *Hvorfor det?*
  (Why is that?)

- discourse deictic is used as the subject complement of "*være* (be) and *blive* (become) in answers (or in coordinated successive clauses):

  (3) a. **A:** *Blev du færdig med opgaven?*
  (Did you finish the task?)
  B: *Ja, det blev jeg*
  (lit. Yes, that did I)
  (Yes, I did)

  b. **A:** *Er du syg?*
  (Are you ill?)
  B: *det er jeg*
  (lit. that am I)
  (Yes, I am)

- discourse deictic corefers with a verb phrase when it is used as the object complement of the verb *have* (have), modal verbs and with the verb *gøre* (do), which replaces the lexical verb in the previous clause in cases where the finite verb of the clause is not an auxiliary or a modal:

(4) a. **A:** *har de set lejligheden?*
(have they seen the apartment?)
B: *det har de*
(lit. that have they)
(Yes, they have)

b. **A:** *Skal du også læse filosofi?*
(Are you also going to study philosophy?)
B: *Nej, det skal jeg ikke*
(lit. No, that am I not)
(No, I am not)

c. *Jeg faldt, men det gjorde hun ikke*
(lit. I fell, but that did she not)
(I fell, but she did not)

– discourse deictic co-refers with an infinitive clause:

(5) *At ryge er farligt og det er også dyrt*
(Smoking is dangerous and it is also expensive)

– discourse deictic corefers with a clause in constructions with attitude verbs and other verbs which take clausal complements, such as *tro* (believe), *vide* (know), *sige* (say) and *prøve* (try):

(6) a. **A:** *er du øm her ved livmoderhalsen*
(does it hurt here by your cervix uteri)
B: *nej ... det tror jeg nu ikke*
(lit. no ... that think I not)
(no.. I don't think so)
SL

b. **A:** *du kan ligeså godt gå fra så tidligt som muligt selvfølgelig*
(you can just as well go on leave as early as possible of course)
B: *ja, det synes jeg*
(lit. yes that think I)
(yes, I think so)
SL

– discourse deictic refers to more clauses, or to something that can be vaguely inferred from the previous discourse (vague anaphora) as it is the case in the following example:

(7) **A:** *nu skal vi jo have ?(lille)? drengen til... i skole her til august jo*
(now we must have the ?(little)? boy in school here in august)
B: *ja*
(yes)
A: *så skal han starte på den der Kreb- eller* (ler)
(then he has to begin in that Kreb- or) (laughs)[3]
B: *skal han det*
(lit. has he that)
(has he?)
A: *ja... han skal,* **det** *vil jeg sgu' godt give ham*
(lit. yes... he has, **that** will I certainly give him)
(yes... he has, I will certainly give it to him)
SL

[3] Here it is referred to Krebsskolen, a private school in Copenhagen.

In example (7) **det** refers to the fact that speaker **A** wants to pay the school fee to his child and allow him to attend a renowned private school. These facts are not explicitly stated in the conversation.

As in English Danish discourse deictics can refer to one or more verbal phrases, one or more clauses, a discourse segment and something that can be vaguely inferred from the context. Furthermore Danish deictics are used in cases where elliptical constructions are common in English and instead of "do so/do too" constructions. A characteristics of Danish discourse deictics is that they often appear before the main verb, in the place that is usually occupied by the subject, as it can be seen in examples (2)–(7). This position is called "fundamentfelt" (actualisation field) in [3].

3 Identifying Discourse Deictics

Discourse deictics are even more common in Danish than in English, especially in dialogues. For instance annotating the pronominal anaphors in four dialogues from the **SL** collection we found that 216 out of 395 personal and demonstrative pronouns were discourse deictics. Although discourse deictics are so common, only one algorithm has been proposed for resolving them, the ES99-algorithm [6,5]. Eckert and Strube, ES99 henceforth, define the ES99-algorithm for resolving anaphors referring to individual nominals and abstract objects in English telephone conversations. The algorithm contains rules for discriminating among the two types of anaphor based on the predicative contexts in which the anaphors occur. Anaphors classified as referring to non-abstract objects are resolved with a centering-based algorithm [17]. Anaphors recognised as discourse deictics are divided into different types and some of them are then resolved with a specific algorithm. ES99 manually test the approach on selected dialogues and obtain a precision of 63,6 % for discourse deictics and of 66,2% for individual anaphors. The precision for individual anaphors is much lower than that obtained when centering-based resolution algorithms are only applied to anaphors with non-abstract antecedents.

The ES99-algorithm was adapted to Danish with slightly better results than those obtained by ES99, but it was found too simplistic for correctly classifying and resolving different types of discourse deixis [16,15]. Although we agree, we believe that the ES99-strategy of identifying discourse deictics from their contexts is useful to NLP systems and that this part of their algorithm is worth pursuing. The strategy is also in line with the studies of English discourse deictics in [8,1]. Thus we have decided to investigate the contexts in which Danish discourse deictics occur and extend the original ES99 rules with both general and Danish specific rules. Most of the rules are preference rules, thus defeasible. Of the rules we present in the following the first four are simply adaptations to Danish of the ES99-rules and are marked with an asterisk "*". The remaining rules have been identified by us.

Rules for identifying Danish discourse deictics:

1. * constructions where a pronoun is equated with an abstract object, e.g., *x er et forslag* (x is a suggestion)
2. * copula constructions with adjectives which can only be applied to abstract entities, such as *x er sandt* (x is true), *x er usandt* (x is untrue), *x er rigtigt* (x is correct)

3. * arguments of verbs which take S'-complements, e.g., *tro* (believe), *antage* (assume), *mene* (think), *sige* (say)
4. * anaphoric referent in constructions such as *x er fordi du er holdt op med at ryge* (x is because you have stopped smoking) *x er på grund af at du er gravid* (x is because you are pregnant)
5. object of *gøre* (do)
6. subject complement with *være* (be) and *blive* (become) in answers or in coordinated clauses
7. object of *have* (have) if the verb was not used as a main verb in the previous clause
8. object of modal verbs
9. in copula constructions where the adjective can both refer to an individual NP and to an abstract object, such as *x er godt* (x is good), *x er dårligt* (x is bad) the anaphor co-refers with an abstract object if the previous clause contains a raising adjective construction, or constructions where an infinite is the subject
10. in constructions where the anaphors are objects of verbs such as *elske* (love), *hade* (hate), *foretrække* (prefer) the anaphor co-refers with an abstract object if the previous clause contains a raising adjective construction or constructions where an infinite clause is the subject (see rule 9)
11. in constructions of the type *det lyder godt* (it sounds well) *det lyder dårligt* (it sounds bad) *det* corefers with a discourse segment unless the previous utterance/ clause contains a nominal or a verb referring to sounds

Rules 9-11 deal with pronouns that can both have an abstract and a non-abstract referent. Rule 10 is illustrated by the following two examples:

(8) a. *Peter ejede det store røde hus ved købmandsbutikken. Det hadede han.*
(lit. Peter owned the big red house near the grocer's store. It hated he.)
(Peter owned the big red house near the grocer's store. He hated it)

b. *Det er dødsygt at sidde på et vaskeri. Det hader jeg.*
(lit. It is boring to be in a laundry. It hate I)
(It is boring to be in a laundry. I hate it)

In example (8-a) the algorithm chooses *det store røde hus ved købmandsbutikken* (the big red house near the grocer's store) as the antecedent of *det*, while in example (8-b) it chooses *at sidde på et vaskeri* (being in a laundry) instead of *et vaskeri*.

It must be noted that in cases as example (8-a), it is often not possible to determine whether the anaphor refers to an individual NP or an abstract object without a deeper analysis of the discourse. Obviously our simple rules will fail to detect these ambiguities.

4 Evaluation

To test the rules we have randomly chosen four dialogues from the **SL**-collection, and manually marked all the occurrences of third singular person neuter personal and demonstrative pronouns as individual anaphors or discourse deictics. Pleonastic uses of *det* (it) have also been marked as non-anaphoric and have been excluded from the test. The rules for identifying discourse deictics have been manually applied to the unmarked dialogues.

Lines in the dialogues containing the constructions indicated by the rules have been automatically extracted from the tagged dialogues[4] and the results have been manually checked. Verbs taking S'-complements and raising adjectives have been automatically marked using the syntactic encodings of the Danish PAROLE lexicon.[5] The results of the human disambiguation and the rule-based identification have then been compared. The success rate for the discriminating algorithm was of 86,13 %. Cases of failure were especially anaphors occurring in constructions allowing for both an individual NP antecedent and an abstract object antecedent, which are not covered by rules 9, 10 and 11. An example are objects of verbs which usually take a concrete object, but are used metaphorically, such as *sluge* (swallow).

One problem with the test we made is that we applied the algorithm on the same type of dialogue which we used to identify the algorithm's rules. Although we have also looked at discourse deictics in a written corpus to identify the rules, it is possible that there are cases of identifiable deictics which we have not covered.

5 Conclusion and Future Work

In the paper we have proposed rules for the (semi-)automatic identification of Danish discourse deictics on the basis of the contexts they occur in. The idea is taken from [6,5]. The first test of these rules gave good results, but it was made on a subset of the dialogues used to identify the rules. Thus they should be tested on other types of dialogue and on written texts. The discriminating rules should also be supplied with a semantic lexicon containing information about metaphorical uses of verbs and nominals referring to abstract objects. Although we believe that the results of the algorithm would be improved by such a lexicon, it is impossible, in our opinion, to discriminate all cases of anaphors which can both refer to non-abstract and abstract objects without a deep analysis of the context.

In this paper we have not addressed at all the issue of how to resolve the identified discourse deictics. However looking at the contexts in which the anaphors occur also helps to identify the type of semantic object referred to by the anaphors [18,8] and we will investigate this aspect in our future work.

References

1. N. Asher. *Reference to Abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.
2. H. Bergenholtz. DK87-DK90: Dansk korpus med almensproglige tekster. In M. Kunø and Erik Larsen, editors, *3. Møde om Udforskning af Dansk Sprog*, 1990.
3. P. Diderichsen. *Elementær Dansk Grammatik*. Gyldendal, Copenhagen, 1957[1946].
4. D. Duncker and J. Hermann. Patientord og lægeord - særord eller fællesord? *Månedsskrift for Praktisk Lægegerning - Tidsskrift for Praktiserende Lægers Efteruddannelse*, pages 1019–1030, 1996.

[4] The dialogues have been tagged with the Brill tagger and a modified set of the PAROLE-tags described in [10].

[5] The general PAROLE specifications are in [7]. The encoding of Danish verbs is described in [14].

5. M. Eckert and M. Strube. Dialogue Acts, Synchronising Units and Anaphora Resolution. In J. van Kuppevelt, N. van Leusen, R. van Rooy, and H. Zeevat, editors, *Amstelogue'99 Proceedings - Workshop on the Semantics and Pragmatics of Dialogue*, 1999.
6. M. Eckert and M. Strube. Resolving Discourse Deictic Anaphora in Dialogues. In *Proceedings of the EACL-99*, pages 37–44, 1999.
7. N. Calzolari (ed.). Parole linguistic resources: Technical specifications overview. Wp4: Deliverable n.4, MLAP PAROLE, 1996.
8. K. Fraurud. *Processing Noun Phrases in Natural Discourse*. Department of Linguistics - Stockholm University, 1992.
9. F. Gregersen and I. L. Pedersen, editors. *The Copenhagen study in urban sociolinguistics*. Reitzel, 1991.
10. D. Haltrup Hansen. Træning og brug af brill-taggeren pĺ danske tekster. Ontoquery, Center for Sprogteknologi, 1999.
11. P. J. Henrichsen. Peeking Into the Danish Living Room. In CST, editor, *Nodalida '98 Proceedings - The 11th Nordic Conference on Computational Linguistics*. Center for Sprogteknologi and Department of General and Applied Linguistics-University of Copenhagen, 1998.
12. J. Hermann. Understandings between doctors and patients - some methodological issues. In *Proceedings ofm the Conference on Medical Interaction, Oct. 18-20*, Odense, 2000. University of Southern Denmark.
13. S.C. Levinson. Pragmatics and the grammar of anaphora: a partial pragmatic reduction of Binding and Control Phenomena. *Journal of Linguistics*, 23(2):379–434, 1987.
14. C. Navarretta. Encoding Danish Verbs in the PAROLE Model. In R. Mitkov, N. Nicolov, and N. Nikolov, editors, *Proceedings of RANLP'97.Recent Advances in Natural Language Processing*, pages 359–363, Tzigov Chark, Bulgaria, 1997.
15. C. Navarretta. Abstract Anaphora Resolution in Danish. In L. Dybkjær, K. Hasida, and D. Traum, editors, *Proceedings of 1st SIGdial Workshop on Discourse and Dialogue - Held in conjunction with The 38th Annual Meeting of the ACL*, pages 56–65, Hong Kong, 2000.
16. C. Navarretta. Centering-based Anaphora Resolution in Danish Dialogues. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of TSD 2000*, pages 345–350, Brno, Czech Republic, 2000.
17. M. Strube. Never Look Back: An Alternative to Centering. In *Proceedings of the 36th Meeting of the ACL*, volume II, pages 1251–1257, Montreal, Quebec, Canada, 1998. Université de Montréal.
18. B. L. Webber. Structure and Ostension in the Interpretation of Discourse Deixis. *Natural Language and Cognitive Processes*, 6(2):107–135, January 1991.