The DAD parallel corpora and their uses

Costanza Navarretta

University of Copenhagen
Njalsgade 140, build. 25
DK- 2300 Copenhagen S
E-mail: costanza@hum.ku.dk

Abstract

This paper deals with the uses of the annotations of third person singular neuter pronouns in the DAD parallel and comparable corpora of Danish and Italian texts and spoken data. The annotations contain information about the functions of these pronouns and their uses as abstract anaphora. Abstract anaphora have constructions such as verbal phrases, clauses and discourse segments as antecedents and refer to abstract objects comprising events, situations and propositions. The analysis of the annotated data shows the language specific characteristics of abstract anaphora in the two languages compared with the uses of abstract anaphora in English. Finally, the paper presents machine learning experiments run on the annotated data in order to identify the functions of third person singular neuter personal pronouns and neuter demonstrative pronouns. The results of these experiments vary from corpus to corpus. However, they are all comparable with the results obtained in similar tasks in other languages. This is very promising because the experiments have been run on both written and spoken data using a classification of the pronominal functions which is much more fine-grained than the classifications used in other studies.

Introduction

In this paper we present an analysis of the uses of abstract pronominal anaphora (abstract anaphora henceforth) annotated in the DAD project (www.cst.dk/dad) and we describe machine learning experiments run on these data. Abstract anaphora refer in the paper to those anaphoric pronouns whose antecedents are predicates in copula constructions, verbal phrases, clauses or discourse segments and whose referents are abstract objects such as properties, events, facts and propositions. In English these pronouns comprise the personal pronoun *it* and the demonstrative pronouns *this* and *that*. We distinguish abstract anaphora from individual anaphora, which have nominal phrase antecedents[1].

The main goal behind the annotation of reference in the DAD project has been to provide annotated data to be used in the study and treatment of abstract anaphora in Danish and Italian written and spoken corpora (Navarretta and Olsen, 2008). This goal is similar to that behind other (co)reference annotation initiatives in other languages, see i.a. (Recasens and Martí, 2010; Dipper and Zinsmeister, 2009; Poesio and Artstein, 2008).

The DAD corpora consist of parallel and comparable corpora of Danish and Italian texts and spoken data which are annotated with information about third person singular neuter personal pronouns and neuter demonstrative pronouns, their functions and their abstract anaphoric uses.

Empirical studies of English, among others (Byron and Allen, 1998; Gundel et al., 2003; Gundel et al., 2004; Gundel et al., 2005, Hedberg et al., 2007) confirm Webber's (1988) observation that, in English, personal pronouns cannot often refer to abstract entities if the antecedent is a clause because the clause is not accessible to the pronoun. Hegarty (2003), Gundel et al. (2005) and Hedberg et al. (2007) explain the frequent use of demonstrative pronouns to refer to clausal antecedents in terms of the *Givenness Hierarchy* (Gundel et al., 1993). Entities introduced in discourse by clauses are only activated in the cognitive status of the addressee, while entities introduced in discourse by verbal phrases are more often in focus and can therefore be referred to by the personal pronoun *it*. Hegarty (2003) and Gundel et al. (2005) also point out that the referents of demonstrative pronouns often are facts, situations and events because clauses refer to these types of entity.

The studies of English abstract anaphora, together with English annotated corpora, have been used in algorithms for resolving *it*, *this* and *that* (Eckert and Strube, 2001; Byron, 2002; Strube and Müller, 2003; Müller, 2007). Studies of abstract anaphora in other languages, among others (Fraurud, 1992; Borthen et al., 1997; Kaiser, 2000; Navarretta, 2002; 2007), indicate that there are language specific characteristics of abstract anaphora which are not captured by the English studies of abstract anaphora. Thus, resolution algorithms developed for resolving English abstract anaphora cannot account for all uses of abstract anaphora in other languages (Navarretta, 2002; Navarretta, 2004).

The paper is organised as follows. In section 2 we account for abstract anaphora in Danish and Italian and describe the DAD data. In section 3 we analyse and discuss the annotated data, while in section 4 and 5 we describe machine learning experiments run on the Danish and Italian data, respectively. Finally, in section 6 we conclude and present future work.

[1] Abstract nouns also refer to abstract objects. However, they are not included in the present study which focuses on pronominal reference.

2. The data

2.1 The pronouns

In Danish texts abstract anaphora comprise the ambiguous pronoun *det* (it/this/that) and the demonstrative pronoun *dette* (this). In spoken language they include the unstressed personal pronoun *det* (it) and the stressed demonstrative pronouns *d’et*[2] (this/that), *d’et h’er* (this) and *d’et d’er* (that). The demonstrative pronoun *dette* is only seldom used in spoken Danish.

In Italian the personal pronouns *lo*, *ne* and *ci* (it non-subject) and the demonstrative pronouns *questo* (this) *quello* (that) and *ciò* (this/that) can be abstract anaphors. The pronouns *lo*, *ne* and *ci* occur both as clitic forms and independent pronouns. Being Italian a subject PRO-drop language, third-person singular verbal forms with implicit subject pronouns (which we call *zero anaphora*) are also annotated.

2.2 The corpora

The DAD corpora consist of a number of subcorpora collected by various research groups: 1. transcriptions of the Danish version of the MAPTASK dialogues comprising 52,145 running words and monologues consisting of 21,224 words (Grønnum, 2006); 2. transcriptions of the AVIP[3] corpus, the Italian version of the MAPTASK corpus, comprising 70,054 words; 3. transcriptions of multiparty spontaneous dialogue extracts from the Danish LANCHART corpus (Gregersen, 2007) comprising 24,112 running words; 4. transcriptions of TV-interviews (2,192 words); 5. three Pirandello’s (1922) stories consisting of 11,139 words and their translations to Danish (11,280 words); 6. articles from an Italian financial newspaper, *Il Sole 24 Ore*, consisting of 13,964 words; 7. Danish and Italian parallel EU texts containing 24,389 and 25,303 words, respectively; 8. Danish texts belonging to the the juridical domain (11,600 words); 9. extracts of newspaper and journal articles, novels and reports (12,570 words) from the Danish general language PAROLE corpus (Keson and Norling-Christensen, 1998). The Pirandello stories, the Italian financial newspaper and the EU texts were collected under the MULINCO project (Maegaard et al., 2006).

All written corpora contain PoS and lemma information. Most of the spoken corpora are also annotated with PoS information, but with different tag sets. The texts are marked with structural information such as chapters and paragraphs, while the transcriptions of spoken data contain information about speakers’ turns and timestamps with respect to the audio files[4]. The two DanPASS corpora also contain rich prosodic information (Grønnum, 2006) while in the multiparty dialogues relevant stress information has been annotated in the DAD project.

2.3 The DAD annotation

The annotation was done in the PALinkA tool (Orasan, 2003) and is available in XML format. The annotation schemes used in the project are extensions of the MATE/GNOME scheme (Poesio, 2004). Two slightly different schemes account for the pronominal systems in Danish and Italian. A description of the schemes and measures of inter-coder agreement for the various categories in terms of weighed *kappa* scores (Cohen, 1968) are in (Navarretta and Olsen, 2008; Navarretta, 2009a).

The following types of information are annotated: a) the type of pronoun, e.g. unstressed *det*, stressed *det*, *dette*, clitic *lo*, zero pronoun; b) the pronominal function such as non-referential, deictic, cataphoric, abstract anaphoric; c) the antecedents of the anaphor; d) the syntactic type of the antecedent; e) the semantic type of the referent, e.g. properties, eventualities, facts, propositions and speech acts[5]; f) the anaphoric distance in term of clauses; g) the relation between anaphor and antecedent ("identity" and "non-identity" relations).

Parts of the text corpora also contain (co)reference information for all types of nominal phrases (Navarretta, 2009a).

3. Exploring the annotated data

In table 1 and 2 we show the pronouns and their main functions in the Danish and Italian data, respectively. In the tables we have grouped vague anaphors, deictics, cataphors textual deictics[6] and abandoned pronouns[7] in the class *Other*.

Pronoun	Non referent.	Indiv. anaph.	Abstr. anaph.	Other	Total
Danish Texts					
Det	345	152	130	81	708
Dette	0	23	71	4	98
Total	345	175	201	85	816
Danish Monologues					
unstressed	22	107	27	54	210
Stressed	1	74	10	45	130
Total	23	181	37	99	340
Danish dialogues					
Unstressed	158	483	299	467	1407
Stressed	10	185	204	197	596
Total	168	668	503	664	2003

Table1: Pronominal types and their functions in Danish

[2] The apostrophe indicates that the following vowel is stressed.

[3] ftp://ftp.cirass.unina.it/cirass/avip.

[4] All the transcriptions are in PRAAT TextGrid format (http://www.praat.org) and have been automatically converted into XML before being annotated in the DAD project.

[5] Most of the semantic types are taken from the middle layer of the of hierarchy abstract objects proposed by Asher (1993).

[6] Textual deictics are the pronouns which refer to, but are not co-referential with, a preceding linguistic expression in the co-text (Lyons, 1977:667-668).

[7] A*bandoned* are the pronouns which occur in unfinished and abandoned utterances.

The data in table 1 confirm previous studies (Navarretta, 2002; Navarretta, 2007) indicating that the most frequently used abstract anaphor in Danish texts is the ambiguous pronoun *det* (65% of the abstract anaphors). *Det* is also the most frequently occurring pronoun and it can be used both referentially and non-referentially. When *det* is used as anaphor it is an individual anaphor in 54% of the cases, while it is an abstract anaphor in the remaining 46% of its occurrences. The demonstrative pronoun *dette* is almost always used as anaphor, in one fourth of its occurrences it refers to individual entities and in all the remaining cases it refers to abstract entities.

In the Danish monologues and dialogues the most frequent abstract anaphora is the unstressed *det*. In the dialogues the stressed *det* is used as abstract anaphor in 34% of the cases while it occurs as individual anaphor in 31% of its occurrences. In the monologues it often refers to individual entities (56% of its occurrences) and it only seldom refers to abstract entities (7% of the cases).

In the spoken data there were only two occurrences of the demonstrative pronoun *dette* (this) and in both cases they referred to individual entities.

Pronoun	Non referential	Indiv. anaph.	Abstr. anaph.	Other	Total
Italian Texts					
Zero	34	317	19	22	392
clitic	0	100	2	4	106
personal	0	165	12	4	181
demonstr.	0	16	7	4	27
total	34	598	40	34	706
Italian Dialogues					
Zero	1	26	42	3	72
clitic	0	19	0	2	21
personal	0	128	11	56	195
demonstr.	0	7	3	10	10
total	1	180	56	71	308

Table 2: Pronominal types and their functions in Italian

The data in table 2 indicate that abstract pronominal anaphors in Italian are quite seldom, as also noticed by Navarretta (2007). In fact, abstract reference is in most cases expressed with abstract nouns in Italian. When abstract pronominal anaphors occur in this language, they are often zero pronouns (48% of the abstract anaphors in the texts and 75% of the abstract anaphors in the dialogues). Zero anaphora, according to models of nominal referring expressions (Givón 1976, Ariel, 1988), are the anaphora which refer to the most salient antecedents in discourse. The use of demonstrative pronouns in abstract reference is extremely seldom in Italian and the most frequently used demonstrative abstract anaphora is the pronoun *ciò* (this/that).

In table 3 we show the abstract pronouns and their antecedent types in the Danish corpora. In the table CL stands for clause, DS for discourse segment (more sentences), CPR for predicate in copula construction, VP for verbal phrase and AA for abstract anaphor. The class CL includes all the clausal types which are distinguished in the data comprising categories such as main clauses, subordinate clauses, matrix clauses and complex clauses (Navarretta and Olsen 2009).

Corpus	Antec	Pronoun	Total	Pronoun	Total
Danish Texts	CL	det	72	dette	60
	DS		6		7
	CPR		13		4
	VP		17		3
	AA		5		3
Danish Monol.	CL	unstressed det	22	stressed det	4
	DS		1		0
	VP		1		2
	CPR		85		62
	AA		19		20
Danish Dialog.	CL	unstressed det	165	stressed det	122
	DS		8		3
	VP		52		35
	CPR		208		57
	AA		149		55

Table 3: Danish abstract anaphora and their antecedents

The following can be observed from the data in the table. In Danish texts the ambiguous pronoun *det* is the most frequently used pronoun when the antecedent is a clause or a discourse segment (54% of its uses). In the monologues and in the dialogues the unstressed pronoun is the most frequently used pronoun with clausal antecedents (96% of the occurrences in the monologues and 58% of the occurrences in the dialogues).

The demonstrative pronoun *dette* (this) in texts seems to be used in contexts where the antecedent is not the most expected one, i.e. it is only part of a preceding complex clause and not the whole clause. In these cases the antecedent is the last occurring (sub)clause. In some cases the use of *dette* indicates that the antecedent is a nominal phrase and not a clause, as it would be expected from the context. This use is opposite to that of the English demonstrative pronouns. Table 4 shows the abstract pronouns and their antecedent types in the Italian corpora.

Corpus	Antec	Pronoun	Total	Pronoun	Total
Italian texts	CL	zero	17	ciò	22
	DS		1		2
	CL	lo/ne	10	questo	-
	DS		1		1
	CPR		1		-
Italian dialogues	CL	zero	41	questo	5
	CL	lo	3		
	VP		5		-
	CPR	ci	2		-

Table 4: Italian abstract anaphora and their antecedents

The data in table 4 indicate that, although abstract pronominal reference in Italian is seldom, zero anaphors and personal pronouns (both clitics and independent forms) are very often used when the antecedent is a clause (55% of the cases in the texts and 90% in the dialogues).

The annotations of the semantic types of the referents, which are not included in the tables, indicate that in Danish the unstressed *det* in spoken data and the ambiguous *det* in texts are the most frequently used pronouns with verbal and clausal antecedents if the referents are eventualities, properties and predicates. All pronominal types occur with equal frequency when the referents are facts.

In Italian all pronouns refer to all types of referents, but zero anaphors and personal pronouns are the most frequently used pronouns when the referred entities are classified as facts. Reference to propositions is done in 85 % of the cases by zero anaphors.

Concluding the annotated Danish and Italian data confirm our initial hypothesis that there are differences in the way various pronominal types are used as abstract anaphora in these two languages compared to the corresponding pronominal types in English. Webber (1991) reports that in a corpus of written English 83.4% of the abstract anaphors were the demonstrative pronouns *this* and *that* and only the remaining 15.6% were occurrences of the pronoun *it*. Similar measures are reported by Byron and Allen (1998) for the TRAINS corpus and by Gundel et al. (2005) for the Santa Barbara Corpus of Spoken English.

Thus, demonstrative pronouns are the most frequently occurring abstract anaphors in both spoken and written English corpora. This is certainly not the case in either Danish or Italian.

Furthermore, Danish and Italian demonstrative pronouns do not have clausal antecedents more often than personal pronouns as it is the case in English (Webber, 1988; Hegarty, 2003; Hedberg et al., (2007; Navarretta, 2007). On the contrary, it seems that clauses are often the most salient entities in Danish and, therefore, they often occur as the antecedents of personal pronouns in this language. The same can be said for Italian in the contexts in which abstract reference is expressed with pronouns.

3.1 Discussion

Previous studies of the uses of abstract anaphora in Swedish (Fraurud, 1992) and Danish (Navarretta, 2002) have pointed out that the ambiguous pronoun *det* is the most frequently used abstract anaphor in texts in the two Scandinavian languages. Borthen et al. (1997) analyse some contexts in which the unstressed pronoun *det* occurs with clausal antecedents in Norwegian. They explain these cases by extralinguistic factors.

Although it is clear that many factors contribute to determine salience in discourse, see among other (Hajičová et al., 1990; Kaiser, 2000; Kaiser and Trueswell, 2004; Gundel et al., 2003; Navarretta, 2002; 2005), we believe that the use of various pronominal types in particular contexts in our data is systematic and thus, it should also be accounted for by the languages' different characteristics, such as their pronominal system and syntactic structure, see also (Navarretta, 2008).

Inanimate entities have only one gender in English, while they belong to two different genders in Danish and Italian. However, only neuter pronouns in Danish and masculine pronouns in Italian can be abstract anaphors. Intuitively, abstract pronominal reference must be more restricted in English than in the other two languages and this can in part explain the frequent use of demonstrative pronouns in English to signal an abstract antecedent compared to Danish and Italian.

Constructions such as clefts and left dislocations are much more frequent in Danish than in English. This is why clauses are more often in focus in the former language than in the latter. The observation that syntactic structure, information structure and salience are strictly related is not new, see i.a. (Sgall et al., 1985; Grosz et al., 1995; Gundel et al., 2003; Navarretta, 2002).

Differing from the other two languages, the order of constituents at the sentence level is free in Italian. This can in part account for the frequent use of abstract substantives in this language. In fact, abstract substantives explicitly indicate the semantic type of the referent excluding ambiguities between individual and abstract referents and reducing the search space for candidate antecedents compared to contexts where pronominal abstract anaphors are used. Our data also show that in Italian abstract anaphora are used in unambiguous contexts or in contexts where the abstract reading is the most natural one. This is compatible with the *Givenness Hierarchy* and accounts for the many occurrences of zero anaphors and personal pronouns with clausal antecedents.

4. Machine Learning Experiments on the Danish Data

In (Navarretta, 2009b) we described machine learning experiments run on the DAD Danish corpora in order to recognise the function of third person singular neuter personal pronouns and neuter demonstrative pronouns.

These experiments were inspired by previous work aimed to recognise some of the functions of the pronoun *it*, see i.a. (Evans, 2000; Müller, 2007) and of the Dutch pronoun *het* (it) (Hoste et al 2007).

Differing from these studies, we ran machine learning algorithms on both written and spoken corpora and looked at the functions of both personal and demonstrative pronouns and of unstressed and stressed pronouns. All the experiments were run in Weka (Witten and Frank, 2005) using the contexts in which the pronouns occur. We worked with four datasets: the Danish texts, the DanPASS monologues, the DanPASS two-party dialogues and the multiparty dialogues. In the first experiments we run unsupervised machine algorithms on the datasets. The results of these experiments indicate that unsupervised learning run on datasets of the size of the DAD corpora do not give satisfactory results for the task of recognizing so fine-grained functions of pronouns as those provided in our annotations because too few clusters are identified and correctness is too low.

In a second group of experiments we ran supervised machine learning algorithms on the four datasets. As training data we used the context in which the pronouns occurred, the pronouns and their functions experimenting with n-grams of various sizes. All experiments were tested using ten-fold cross validation. The baseline is provided by the results of the Weka ZeroR classifier that predicts the most frequent attribute value for nominal classes. We tested various classifiers following a strategy proposed by Daeleman et al. (2003). The algorithms which gave the best results on the data are the following: NBTree which generates a decision tree with Naive Bayes classifiers at the leaves, SMO which is an implementation of a support vector machine and K-star, an instance-based classifier which uses an entropy-based distance function.

The baseline and the best results achieved on the Danish data are in table 5 (Navarretta, 2009b). The results are given in terms of precision, recall and F-measure which in Weka are calculated as weighted averages of the results obtained for each class.

Algorithm	Precision	Recall	F-measure
Danish Texts			
Baseline	18.3	42.8	25.7
NBTree	62.3	65.4	62.4
DanPASS Monologues			
Baseline	28.3	53.2	37
SMO	64.3	66.8	64.7
DanPASS Dialogues			
Baseline	15.1	38.8	21.7
SMO	54.5	57.2	55.4
Multiparty Dialogues			
Baseline	9	30	13.8
Kstar	33.4	35.4	32.9

Table 5: Classification results on Danish data

These results show that classification improves the recognition of the pronominal functions on texts, monologues and two-party dialogues with more than 35% with respect to the baseline, while only a 19% improvement was achieved on multi-party dialogues. Not surprisingly, the best results were achieved on the monologues which are the most homogeneous dataset. The obtained results indicate that classifiers can be useful to tag the function of pronouns in texts, in monologues and in some types of dialogues, although the annotation still needs manual correction.

The performance of classification on the multiparty dialogues was not as good as that obtained on the DanPASS dialogues. This can be partly explained by the fact that the former dialogues are spontaneous and less homogeneous than the latter. Furthermore, the annotations did not contain information about adjacency pairs and this type of information is very important when processing multiparty dialogues.

In a third group of experiments we ran classification algorithms on the text dataset to which we had added lemma and PoS information. The purpose of these experiments was to investigate whether these types of linguistic information improve classification. Here we followed the strategy adopted by Hoste et al. (2007). The performance of the classifiers improves when PoS and lemma information are added to the data, but the improvement is not significant.

The classification results obtained on texts, monologues and two-party dialogues are comparable with those reported by i.a. Hoste et al. (2007) for the Dutch pronoun *het*. The recognition of non-referential occurrences of the pronouns in the various datasets is slightly lower than that obtained by Boyd et al. (2006) with a system which recognises occurrences of the non-referential *it* using word patterns and list of weather verbs and idioms.

Our results are promising because we identify more pronominal functions than other researchers. Furthermore, we account for the occurrences of all third person neuter pronouns in both written and spoken data using a more fine-grained classification of the pronominal functions than those used by other researchers. The results of our experiments also indicate that the granularity of the function classification used in the DAD project can be used to train classification algorithms.

5. Machine Learning Experiments on the Italian Data

In the present experiment we have run supervised machine learning experiments on the Italian DAD data. These experiments are similar to those described in the preceding section. The following two datasets have been used: data extracted from the annotated Italian texts and data extracted from the Italian dialogues. As in (Navarretta, 2009b) we experimented with many classifiers and ran the algorithms on n-grams of various sizes. The best results were achieved on texts considering a window of three words preceding and five words following the “pronoun”. By “pronoun” here we mean the independent pronominal forms and the words explicitly containing one or more clitic pronouns or implicitly including a zero pronouns.

The best results were achieved on the dialogue dataset considering two words before and three words after the pronoun. The results of the three best performing algorithms on each dataset and those obtained by ZeroR (the baseline) are shown in table 6.

The improvement of classification with respect to the baseline is of 55.1% for texts and 35.9% for dialogues. These results are significantly better than those obtained on the Danish data although the Italian corpora are smaller than the Danish corpora.

One reason for the better performance of classifiers on the Italian data than on the Danish data is that there are more pronominal types in Italian than in Danish, thus the use of each pronoun is much more restricted in the former language than in the latter one.

Corpus	Algorithm	Precision	Recall	F-measure
Texts	Baseline (ZeroR)	18.5	43	25.8
	SMO	79.8	84.3	81.1
	NBTree	78	83.6	80.4
	Naive Bayes	71.6	76	73.5
Dialogues	Baseline (ZeroR)	25.3	50.3	33.7
	Kstar	68.9	72.4	69.6
	SMO	63.5	69.2	65.7
	NBTree	63	68.5	64.5

Table 6: Classification results on Italian data

The confusion matrices of the best algorithms on both texts and dialogues indicate, not surprisingly, that the most frequently occurring classes are those that are recognised more correctly by the classification algorithms. This is the case for individual anaphora (both implicit and explicit pronouns) and for expletives in texts and for explicit individual anaphors and zero abstract and zero individual anaphors in dialogues.

Also for Italian we have run a second group of experiments with the purpose of investigating whether PoS and lemma information improves the classification of the function of pronouns. In these experiments we have only used the text dataset and have chosen as classifier SMO because it gave the best results on this dataset in the preceding experiments. We also use as baseline the results obtained by SMO in the preceding experiments where only the pronominal contexts, the pronouns and their pronominal functions were used as training data.

The results of the second group of experiments are shown in table 7.

Dataset	Precision	Recall	F-measure
words (baseline)	79.8	84.3	81.1
words+PoS	78.5	83.8	80.3
words+lemma	79.7	84.1	80.1
words+PoS+lemma	78.8	83.9	80.4

Table 7: Classification results on the Italian texts with PoS and lemma information

The table shows a decrease in the classifier's performance when PoS and lemma information are added to the Italian data, although the decrease in performance is not significant. Because we do not know the performance of the PoS tagger and lemmatiser used to annotate the Italian texts, these experiments should be run again after having corrected the PoS and lemma annotation manually.

The results of our machine learning experiments on the Italian data indicate that supervised machine learning can be a useful support in the task of identifying the function of third-person singular pronouns.

6. Conclusions

In the paper we have described the DAD Danish and Italian corpora which contain information about the occurrences of third-person singular neuter personal pronouns and neuter demonstrative pronouns, their functions and their anaphoric uses with particular focus on the occurrences of abstract anaphors. Then we have described the uses of the abstract anaphors in the data which clearly indicate some systematic differences in the way these anaphors are used in Danish and Italian with respect to abstract anaphors in English. We explain some of these differences looking at the three languages' pronominal systems and syntactic characteristics.

Finally, we have described machine learning experiments run on the Danish and Italian data with the purpose of recognising the functions of singular neuter pronouns. The results of the experiments on the Danish data are comparable with the results obtained in English and Dutch for similar tasks and are particularly promising because we work with more types of data and use a more fine-grained classification of the function of the pronouns than those used in the English and Dutch experiments. The results of supervised machine learning applied to the Italian DAD corpora are much better than those obtained on the Danish data. These results can be explained by the fact that there are more types of pronoun in Italian than in Danish, thus the use of each type of pronoun is much more restricted in the former than in the latter. Adding PoS and lemma information on the Danish data improves classification, but the improvement is not significant. Adding the same type of information to the Italian data decreases classification slightly. This is probably due to the performance of the used tagger and lemmatiser.

The annotations provided in the DAD corpora are useful not only to analyse abstract anaphora in the two languages accounted for, but also to apply supervised machine learning to the data.

Presently, we are testing the performance of the classifiers on the DanPASS data including various types of prosodic information in order to investigate whether this information improves the identification of the pronominal functions.

Future work consists in analysing other information annotated in the corpora, such as the relation between type of pronoun, syntactic antecedent and anaphoric distance. We plan also to look at which features should be included in the datasets to extend the machine learning experiments to the resolution of the abstract anaphors in the data.

6. Acknowledgements

The DAD project was funded by the Danish Research Councils in the period 2007-2009. We especially thank our colleague Sussi Olsen, one of the annotators of the Danish data.

7. References

Ariel, M. (1988). Referring and accessibility. *Journal of Linguistics*, 24(1):65–87.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Boyd, A., Gegg-Harrison, W., Byron, D. (2005).

Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, pp. 40–47, Ann Arbor Michigan, June.

Borthen, K., Fretheim, T., Gundel, J. K. (1997). What brings a higher-order entity into focus of attention? Sentential pronouns in English and Norwegian. In R. Mitkov and B. Boguraev (Eds.), *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, 88–93.

Byron, D.K. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of ACL'02*, 80–87.

Cohen, J. (1968). Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bullettin*, 70:213–220.

Byron, D., Allen, J. (1998). Resolving Demonstrative Pronouns in the TRAINS93 corpus. In *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC 2)*, pp. 68–81.

Daelemans, W., Hoste, V., De Meulder, F., Naudts, B. (2003). Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language. In *Proceedings of the 14th European Conference on Machine Learning (ECML-2003)*, pp. 84–95, Cavtat-Dubrovnik, Croatia.

Dipper, S., Zinsmeister, H. (2009). *Annotating Discourse Anaphora* In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), ACL-IJCNLP, pp. 166-169. Singapore.

Eckert, M., Strube, M. (2001). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Evans, R. (2000). A comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal It. In *NLP 2000*, LNCS 1835, pp. 233–240. Springer-Verlag, Berlin Heidelberg.

Fraurud, K. (1992). *Processing Noun Phrases in Natural Discourse*. Department of Linguistics - Stockholm University.

Givón, T. (1976). Topic, Pronoun and Grammatical Agreement. In Charles N. Li, editor, *Subject and Topic*, pp. 149–188. Academic Press.

Grosz, B., Joshi, A. K., Weinstein, S. (1995), Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics* 21(2), 203–225.

Gregersen, F. (2007). The LANCHART Corpus of Spoken Danish, Report from a corpus in progress. In *Current Trends in Research on Spoken Language in the Nordic Countries*, pp. 130–143. Oulu University Press.

Grønnum, N. (2006). DanPASS - A Danish Phonetically Annotated Spontaneous Speech Corpus. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk and D. Tapias (Eds.), *Proceedings of LREC-06*, Genova, Italy, May.

J. K. Gundel, J.K., Hedberg, N., Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Gundel, J.K., N. Hedberg, N., Zacharski, R. (2003). Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12:281–299.

Gundel, J.K., N. Hedberg, N., Zacharski, R. (2004). Demonstrative pronouns in natural discourse. In A. Branco, T. McEnery and R. Mitkov (Eds.), *Proceedings of DAARC-2004- 5th Discourse Anaphora and Anaphora Resolution Colloquium*, pp. 81–86, Furnal, S.Miguel, Portugal. Ediçoes Colibri.

Gundel, J. K., Hedberg, N., Zacharski, R. (2005). Pronouns without NP Antecedents: How do we know when a pronoun is referential. In Antonio Branco, Tony McEnery and Ruslan Mitkov (Eds.), *Anaphora Processing: Linguistic, Cognitive and Computational Modelling.* John Benjamins. 351-364.

Kaiser, E. (2000). Pronouns and demonstratives in Finnish: Indicators of referent salience. In P. Baker, A. Hardie, T. McEnery and A. Siewierska (Eds.), *DAARC 2000*, volume 12 of *University Center for Computer Corpus Research on Language - Technical Series*, 20–27, Lancaster, UK.

Kaiser, E., Trueswell, J. (2004). The referential properties of Dutch pronouns and demonstratives: Is salience enough? In Cècile Meier and Matthias Weisgerber, editors, *Proceedings of the Conference Sub8 Sinn und Bedeutung*, Arbeitspapier Nr. 177, pp. 137–149, FB Sprachwissenschaft. Konstanz, Germany. Universität Konstanz.

Keson, B., Norling-Christensen, O. (1998). PAROLE-DK. Technical report, Det Danske Sprog- og Litteraturselskab.

Lyons, J. (1977). *Semantics*, volume I-II. Cambridge University Press.

Hajičová, E., Kuboň, P., Kuboň, V. (1990). Hierarchy of Salience and Discourse Analysis and Production. In H. Karlgren (Ed.), *Proceedings of COLING'90*, volume III, 144–148, Helsinki.

Hedberg, N. Gundel, J.K., Zacharski, R. (2007). Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In A. Branco, T. McEnery, R. Mitkov and F. Silva (Eds.), *In Proceedings of DAARC-2007 - 6th Discourse Anaphora and Anaphora Resolution Colloquium*, pp. 31–36, Lagos, Portugal, March. Centro de Linguistica da Universidade do Porto.

Hegarty, M. (2003). Semantic types of abstract entities. *Lingua*, 113:891–927.

Hoste, V., Hendrickx, I., Daelemans, W. (2007). Disambiguation of the Neuter Pronoun and Its Effect on Pronominal Coreference Resolution. In V. Matousek and P. Mautner (Eds.) *TSD 2007*, volume 4629 of *Lecture Notes in Computer Science*, 48–55. Springer.

Maegaard, B., Offersgaard, L., Henriksen, L., Jansen, H., Lepetit, X., Navarretta, C., Povlsen, C. (2006). The MULINCO corpus and corpus platform. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk and D. Tapias (Eds.), *Proceedings of LREC-06*, pp. 2148–2153, Genova, Italy, May.

Müller, C (2007). Resolving it, this and that in unrestricted multi-party dialog. In *Proceedings of ACL-2007*, pp. 816–823, Prague.

Navarretta, C., Olsen, S. (2008). Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-2008*, Marrakesh, Morocco, ELRA, pp. 2046-2052.

Navarretta,C., Olsen,S. (2009). *The annotation of pronominal abstract anaphora in Danish texts and dialogues*. DAD report 1. Centre for Language Technology, University of Copenhagen. January, p.20.

Navarretta, C. (2002). The Use and Resolution of Intersentential Pronominal Anaphora in Danish Discourse. Ph.D. thesis, University of Copenhagen.

Navarretta, C. (2004). Resolving Individual and Abstract Anaphora in Texts and Dialogues. In *Proceedings of COLING-2004*, Geneva, Switzerland, 233-239.

Navarretta, C. (2005). Combining Information Structure and Centering-based Models of Salience for Resolving Danish Intersentential Pronominal Anaphora. In: A. Branco, T. McEnery and R. Mitkov (Eds.), *Anaphora Processing. Linguistic, cognitive and computational modeling*. John Benjamins Publishing Company, 329-350.

Navarretta, C. (2007). A contrastive analysis of abstract anaphora in Danish, English and Italian. In: A. Branco, T. McEnery, R. Mitkov and F. Silva (Eds.), *Proceedings of DAARC 2007 - 6th Discourse Anaphora and Anaphora Resolution Colloquium*, March, Centro de Linguistica da Universidade do Porto, 103-109.

Navarretta. C. (2008). Pronominal types and abstract reference in the Danish and Italian DAD Corpora. In C. Johansson (Ed.), *Proceedings of the Second Workshop on Anaphora Resolution (WAR II)*. NEALT Proceedings Series, Vol. 2: 63-71.

Navarretta, C. (2009a) Co-referential chains and discourse topic shifts in parallel and comparable corpora. *Revista de Procesamiento de Lenguaje Natural*, La Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), 42:105-112.

Navarretta, C. (2009b). Automatic recognition of the function of third-person singular pronouns in texts and spoken data. In: S. Lalitha Devi, A. Branco and R. Mitkov (Eds.), *Anaphora Processing and Applications. 7th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2009 Goa, India, November 5-6, 2009 Proceedings* LNAI 5847. pp. 15-28. Springer Verlag Berlin/Heidelberg.

Orasan, C.(2003). PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pp. 39–43, Sapporo.

Pirandello, L. (1922). *Novelle per un anno*. Giunti.

Poesio, M. (2004). The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In Proceedings of the 5th SIGDIAL Workshop, pp. 154–162, Boston.

Poesio, M., Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC-2008*, Marrakech, Morocco, pp. 1170-1174.

Recasens, M., Martí, M.A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*. DOI 10.1007/s10579-009-9108-x.

Sgall, P., Hajičová, E., , Panevová, J.(1986), *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*, Reidel, Dordrecht.

Strube, M., Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the ACL'03*, pp. 168–175.

Prince, E.F. (1981). Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pp. 223–255. Academic Press.

Webber, B.L (1988). Discourse deixis and discourse processing. Technical report, University of Pennsylvania.

Webber, B.L. (1991). Structure and Ostension in the Interpretation of Discourse Deixis.. 6:107-135.

Witten, I.H., Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition.