Automatic Recognition of the Function of Singular Neuter Pronouns in Texts and Spoken Data

Costanza Navarretta*

Centre for Language Technology
University of Copenhagen
Njalsgade 140-142, build. 25
2300 Copenhagen S
Denmark
costanza@hum.ku.dk
http://cst.dk/costanza

Abstract. In this paper we describe the results of unsupervised (clustering) and supervised (classification) learning experiments with the purpose of recognising the function of singular neuter pronouns in Danish corpora of written and spoken language. Danish singular neuter pronouns comprise personal and demonstrative pronouns. They are very frequent and have many functions such as non-referential, cataphoric, deictic and anaphoric. The antecedents of discourse anaphoric singular neuter pronouns can be nominal phrases of different gender and number, verbal phrases, adjectival phrases, clauses or discourse segments of different size and they can refer to individual and abstract entities. Danish neuter pronouns occur in more constructions and have different distributions than the corresponding English pronouns *it*, *this* and *that*.
The results of the classification experiments show a significant improvement of the performance with respect to the baseline in all types of data. The best results were obtained on text data, while the worst results were achieved on free-conversational, multi-party dialogues.

Key words: singular neuter pronouns, pronominal functions, machine learning, individual and abstract anaphora, text and spoken corpora, annotation

1 Introduction

In this paper we describe the results of unsupervised (clustering) and supervised (classification) learning experiments with the purpose of recognising the function of singular neuter personal and demonstrative pronouns (sn-pronouns henceforth) in Danish corpora of written and spoken language. Therefore, we will relate our work to relevant work done on English and Dutch data. Danish

* Sussi Olsen, Hanne Fersøe and Patrizia Paggio

sn-pronouns are very frequent and have many functions such as non-referential (expletive henceforth), cataphoric, deictic and anaphoric. The antecedents of discourse anaphoric sn-pronouns can be nominal phrases of different gender and number, verbal phrases, adjectival phrases, clauses or discourse segments of different size and they can refer to individual and abstract entities (individual and abstract anaphors, respectively). Danish sn-pronouns occur in more constructions and have different distributions than the corresponding English pronouns *it*, *this* and *that*.

The first step towards the resolution of the anaphoric occurrences of sn-pronouns is their identification and classification with respect to their type of antecedent, see also [7], and this is the subject of the paper. The main goals of our work have been the following: i) to test how well unsupervised and supervised learning algorithms identify the function of Danish sn-pronouns in texts and spoken data; ii) to individuate the information which is most useful to this task; iii) to evaluate the function classification provided in the annotated corpora which we used.

We start by discussing related work in section 2; then we present the data which we have used in section 3; we describe our machine learning experiments and discuss the obtained results in section 4; finally we conclude and present work still to be done in section 5.

2 Related Work

To our knowledge there is no previous work to automatically recognise the function of Danish sn-pronouns. Some algorithms to resolve English pronominal anaphora presuppose pre-editing of the data to allow for the exclusion of non-referential and cataphoric occurrences of pronouns, other algorithms include the identification of some of the pronominal functions[1].

When full-parsing of data is not possible or desirable, filtering mechanisms and selectional preferences are applied to the data to identify the main functions of pronominal occurrences and exclude some of them from the resolution algorithms, see among others [8, 18, 23, 19].

The resolution of the English pronouns *it*, *this* and *that* in English dialogues has been addressed in [7, 3, 25, 19]. Eckert and Strube's algorithm [7] relies on complex knowledge about language and discourse structure and identifies individual and abstract occurrences of third-person singular pronouns in English on the basis of the context in which the pronouns occur and of their type (personal or demonstrative). The algorithm has only been tested manually and non-anaphoric occurrences of the pronouns were excluded from the test. The same method has been partly adapted and incorporated in an algorithm for resolving Danish discourse pronominal anaphora [20, 21]. Also the Danish algorithm has only been tested manually, relies on many knowledge sources and only accounts for pronominal anaphoric occurrences. Byron's PHORA-algorithm [3] resolves the

[1] A comparison of the most known resolution algorithms including information on how much pre-editing and pre-processing they require can be found in [17].

occurrences of *it*, *this* and *that* in domain-specific dialogues. It is implemented and relies on semantic knowledge and a speech act model. An other implemented algorithm for resolving the same English pronouns is described in [25]. This algorithm relies on various types of linguistic information extracted from the Penn Treebank. Finally a machine learning approach for identifying and resolving third-person singular pronouns in English is proposed in [19]. The algorithm has been trained and tested on five dialogues, which were annotated for this task, and relies exclusively on the corpus annotation. The algorithm is exposed to all occurrences of *it*, but the non-anaphoric occurrences were pre-annotated in the data in order to trigger all types of negative preferences which allowed the system to sort them out. The results of this algorithm are much lower than those obtained by the algorithms relying on complex linguistic and discourse structure knowledge.

A machine learning approach for recognising non-referential occurrences of the English *it* in a text corpus is presented in [1]. In this approach some of the rules implemented in rule-based systems are generalised via word patterns which are added to the system as features. The system also uses external knowledge sources in the form of two word lists containing weather verbs and idioms. The system achieved the best results using 25 features (precision was 82% and recall 72% on the given corpus).

The classification of referential and non-referential uses of the Dutch pronoun *het* (it) in two text corpora is described in [12]. The classification comprises the following uses of this pronoun: individual and abstract anaphoric, non-referential, anticipatory subject and anticipatory object. The reported results of the classification give an improvement of approx. 30% for all distinctions with respect to the baseline (the most frequent class). In [12] the authors also measure the effects of the classification on a machine learning based co-reference resolution system.

Our research is inspired by most of these approaches, especially the work described in [7, 20, 1, 12]. The novelty of our approach, apart from the language which we investigate, consists in the following:

- we use both texts and spoken data of various type;
- we deal with personal and demonstrative pronouns as well as weak and strong pronouns in spoken data (prosodic information about stress is included);
- we rely on a very fine-grained classification of the functions of Danish sn-pronouns which covers all occurrences of these pronouns in both texts and spoken data.

In these experiments we only use n-grams of words and, on texts, very basic linguistic information. We start from the raw data (no annotation at all) and investigate to which extent machine learning algorithms (first unsupervised than supervised) can be useful to identify the function of sn-pronominal occurrences. In the supervised experiments we first consider n-grams of words and the classification of sn-pronouns in the data, then we test the learning algorithms adding to the words in the texts lemma and PoS information. In this we follow the

strategy proposed by [6] which consists in testing various machine learning algorithms and types of linguistic information to find the most appropriate datasets and algorithms to resolve NLP tasks.

3 The data

In written Danish sn-pronouns comprise the pronoun *det* (it/this/that), which is ambiguous with respect to its pronominal type, and the demonstrative pronoun *dette* (this). In spoken language they comprise the unstressed personal pronoun *det* (it), the stressed demonstrative pronouns *d'et* (this/that), *d'et her* (this) and *d'et der* (that). The stressed demonstrative pronoun *d'ette* occurs very seldom in spoken language (there were only two occurrences of it in our data and they both referred to an individual entity).

3.1 The corpora

The corpora we use have been collected and annotated by several researcher groups for different purposes. Thus they are very heterogeneous.

The written corpora comprise general language texts [14], legal texts and literary texts [16]. They consist of 86,832 running words. The spoken language corpora comprise transcriptions of monologues and two-party dialogues from the DanPASS corpus [10], which is a Danish version of the MAPTASK corpus, multi-party verbose dialogues from the LANCHART corpus [9] and interviews from Danish television (LANCHART+TV henceforth). The monologues consist of 23,957 running words; the DanPASS dialogues contain 33,971 words and the LANCHART+TV consists of 26,304 words.

3.2 The annotation

All texts contain automatically acquired PoS-tag and lemma information. Most of the spoken corpora are also PoS-tagged, but with different tagsets. The texts contain structural information such as chapters, sections and paragraphs, while the transcriptions of spoken language contain information about speakers' turns and timestamps with respect to the audio files[2]. All sn-pronouns in the spoken data are marked with stress information. The DanPASS data also contain rich prosodic information.

In all corpora sn-pronouns and their functions are marked. (Co)reference chains of the anaphoric sn-pronouns are also annotated together with other linguistically relevant information, such as the syntactic type of the antecedent, the semantic type of the referent and the referential relation type, see [22].

The corpora are available in the XML-format produced by the PALinkA annotation tool [24]. The classification of the function of sn-pronouns provided in the data is very fine-grained. It comprises the following classes:

[2] All the transcriptions were provided in the PRAAT TextGrid format (http://www.praat.org).

- expletive (all non-referential uses);
- cataphoric (the pronoun precedes the linguistic expression necessary to its interpretation);
- deictic (the pronoun refers to something in the physical word);
- individual anaphoric;
- individual vague anaphoric (the individual antecedents are implicit in discourse);
- abstract anaphoric;
- abstract vague anaphoric (the abstract antecedents are implicit in discourse);
- textual deictic (the anaphors refer to, but are not co-referential with, preceding linguistic expressions [15]);
- abandoned (the pronouns occur in unfinished and abandoned utterances[3]).

80% of the corpora were annotated independently by two expert annotators and then the two annotations were compared. The remaining 20% of the data were only coded by one annotator and revised by the other. In case of disagreement the two annotators decided together which annotation to adopt. In difficult cases a third linguist was consulted to choose an annotation. The annotators could listen to the audio files when coding the spoken data.

Inter-coder agreement was measured in terms of *kappa* scores [5, 4] on the first subset of the annotated data (most of the text corpora and the DanPASS dialogues).

Table 1 shows the *kappa*-scores for the most frequent pronominal functions as they are reported in [22].

function	text corpora	DanPASS dialogues
expletive	0.83	0.77
cataphor	0.73	0.72
individual	0.90	0.88
individual vague	0.92	0.92
abstract	0.89	0.84
abstract vague	0.8	0.84
textual deictic	0.91	0.89

Table 1. Intercoder agreement as *kappa* scores

4 The Experiments

The learning experiments have been run in the WEKA system [26] which permits testing and comparing a variety of algorithms. It also provides an interface with which to explore the data and the learning results. We ran the experiments on four datasets automatically extracted from the annotated corpora and translated

[3] These are also called disfluencies in the literature.

into the *arff*-format required by WEKA. The four datasets we distinguish in our experiments are the following:

1. the texts
2. the DANPASS monologues
3. the DANPASS dialogues
4. the LANCHART+TV dialogues.

The sn-pronouns and their functions in the four datasets are given in table 2. The following abbreviations are used in the table: *Expl* for expletive, *IndAna* for individual anaphor, *AbsAana* for abstract anaphor ,*VagIA* for vague individual anaphor, *VagAA* for vague abstract anaphor, *Catap* for cataphor, *Deict* for deictic, *TDeic* for textual-deictic, *Aband* for abandoned.

Pronoun	Expl	IndAna	AbsAna	VagIA	VagAA	Catap	Deict	Tdeic	Aband	Total
Texts										
det	345	152	130	8	10	58	1	4	0	708
dette	0	23	71	0	4	0	0	0	0	98
all	345	175	201	8	14	58	1	4	0	816
DANPASS Monologues										
unstressed	22	107	27	14	1	14	0	0	25	210
stressed	1	74	10	8	13	11	1	0	12	130
all	23	181	37	22	14	25	1	0	37	340
DANPASS dialogues										
unstressed	34	177	100	25	5	17	0	4	72	434
stressed	10	121	111	22	7	22	7	3	31	334
all	44	298	211	47	12	39	7	7	103	768
LANCHART+TV										
unstressed	124	301	199	56	16	128	8	5	138	975
stressed	0	69	93	10	7	32	1	2	46	260
all	124	370	292	66	23	160	9	7	184	1235

Table 2. Sn-pronouns and their functions in the data

4.1 Clustering experiments

Clustering was run on the raw data, but the pronominal function information in the annotated data was used to evaluate the obtained clusters. The best results in terms of the highest number of recognised clusters and "correctness"[4] were

[4] Correctness is calculated by WEKA in the test phase by assigning to each cluster the pronominal function which in the evaluation data is attributed to the largest number of items in that cluster. The function assignment is optimised with respect to the recognised clusters. A *no-class* tag is assigned to clusters whose items have functions which have already been assigned to other clusters. Finally, correctness is calculated for the clusters which have been assigned a function.

achieved by the WEKA EM (Expectation Maximisation). Clustering was tested on n-grams of varying size. The best results on the text data were achieved with a window of one word preceding and two words following the sn-pronouns. Five clusters were returned and they were bound to individual anaphor, expletive, cataphor, abstract anaphor and no-class. Correctness was 37.5 %. The best results on the DANPASS monologues were obtained using a window of 2 words preceding and 3 words following the sn-pronouns. Five clusters were recognised which were bound to the functions individual anaphor, abandoned, vague abstract anaphor, expletive and abstract anaphor. Correctness was 41.5%. On the DANPASS dialogues the best results were obtained with a window of 2 words preceding and following the sn-pronouns. The pronouns from the DANPASS dialogue data were grouped into 4 clusters (abandoned, individual anaphor, abstract vague and cataphor) and correctness was 43.5%. On the LANCHART+TV data the best results were achieved with a window of two words preceding and four words following the sn-pronouns. The algorithm returned 3 clusters connected to the functions individual anaphor, abstract anaphor and expletive. Correctness was 29.5 %.

The fact that clustering gives the best results on the text data confirms that it is harder to process transcriptions of spoken data than written data because other information available in spoken language is not included in the transcriptions.

From the experiments we can conclude that unsupervised learning on datasets of the size we are working with does not provide satisfactory results for the task of recognising such fine-grained functions of sn-pronouns (too few clusters were identified and correctness was too low).

4.2 Classification on words

In the classification experiments we trained several classifiers on data extracted from the corpora. The pronominal function annotated in the corpora was used both for training and testing the classifiers. We started running various classifiers on n-grams as in the clustering experiments, then we run them on the data enriched with various types of information. The latter experiments have only been run on text data. In all cases the results were tested using 10-fold cross-validation. As baseline in our evaluation we used the results provided by the WEKA ZeroR class that predicts the most frequent attribute value for a nominal class (accuracy is the frequency of the most used category). The Weka algorithms which we have tested are: Naive Bayes, SMO, IBK, LBR, KStar, NBTree, LADTree, and Rotation Forest. The algorithms were tested on windows of various sizes (going from the largest one: 3 words before and 5 words after the sn-pronouns to the smallest one: 1 word before and 2 words after the sn-pronouns).

For texts the best results were achieved by the WEKA NBTree class (it generates a decision tree with Naive Bayes classifiers at the leaves) and the dataset comprised three words before and five words after the sn-pronouns. For monologues the best results were obtained by the SMO class (Sequential Minimal

Optimization) run on a window of one word before and three words after the sn-pronouns. For all dialogues the best results were achieved using a window of 2 words preceding and 3 words following the sn-pronouns. On the DanPASS dialogue data the algorithm that gave the best results was the WEKA SMO class, while for the LANCHART+TV data the best results were obtained by the KStar[5] class. The results of the classification algorithms in terms of Precision, Recall and F-measure are in table 3. The table shows the baseline and the three best results obtained for each datasets by various algorithms. Figures 1, 2, 3 and 4

Algorithm	Precision	Recall	F-measure
Texts			
Baseline	18.3	42.8	25.7
NBTree	62.3	65.4	62.4
NaiveBayes	61.1	64.4	61.4
RotationForest	60.7	63.5	60.4
Monologues			
Baseline	28.3	53.2	37
SMO	64.3	66.8	64.7
KStar	63.2	66.5	61.3
IBK	59.6	63.5	60.9
DDialogues			
Baseline	15.1	38.8	21.7
SMO	54.5	57.2	55.4
NaiveBayes	52.9	56.6	53.2
RotationForest	49.9	53.4	50
LDialogues			
Baseline	9	30	13.8
KStar	33.4	35.4	32.9
NBTree	32.9	36.6	32.8
SMO	32.3	33.6	32.7

Table 3. Classification results: words and pronominal function

show the confusion matrices produced by the algorithms that performed best on each of the four datasets. From the confusion matrices it is evident that the performance of classification is bound to the frequency of the various types of item in the data: occurrences of pronouns with frequently used functions are better classified than occurrences of pronouns with seldomly occurring functions such as textual deictic, deictic and, in some datasets, vague anaphor. Thus the confusion matrices reflect the differences in the distribution of the pronominal functions in the various datasets.

From the confusion matrices it can also be seen that cataphors, and individual and abstract anaphors are often confused with expletives. Distinguishing between cataphors and expletives was also problematic for the annotators espe-

[5] K-star is an instance-based classifier which uses an entropy-based distance function.

a	b	c	d	e	f	g	h	<-- classified as
316	4	9	16	0	0	0	0	a = expletive
35	11	8	4	0	0	0	0	b = cataphor
48	1	78	46	0	2	0	0	c = indiv
49	5	28	119	0	0	0	0	d = abstr-ana
7	1	2	4	0	0	0	0	e = abstr-vague
2	0	0	3	0	3	0	0	f = indiv-vague
0	0	1	0	0	0	0	0	g = deictic
0	0	1	3	0	0	0	0	h = textual-deictic

Fig. 1. Confusion matrix for texts

a	b	c	d	e	f	g	h	<-- classified as
13	0	0	0	0	0	0	10	a = explet
0	173	4	2	0	2	0	0	b = indiv
0	17	6	0	0	1	1	0	c = cataphor
0	15	1	6	0	0	0	0	d = indiv-vague
0	0	0	0	0	0	1	0	e = deictic
0	7	2	2	0	26	0	0	f = abstr-ana
0	4	2	1	0	0	7	0	g = abstr-vague
6	0	0	0	0	0	0	31	h = abandoned

Fig. 2. Confusion matrix for monologues

a	b	c	d	e	f	g	h	<-- classified as
6	12	1	1	0	1	0	2	a = explet
7	156	4	2	0	2	0	10	b = indiv
1	15	6	0	0	1	1	1	c = cataphor
1	13	1	5	0	0	0	2	d = indiv-vague
0	0	0	0	0	0	1	0	e = deictic
0	7	2	2	0	26	0	0	f = abstr-ana
0	4	2	1	0	0	7	0	g = abstr-vague
2	10	1	1	0	2	0	21	h = abandoned

Fig. 3. Confusion matrix for DanPASS dialogues

a	b	c	d	e	f	g	h	i	<-- classified as
21	0	13	20	2	63	0	0	5	a = explet
2	1	8	0	0	10	0	0	2	b = abstr-vague
6	3	124	17	6	109	1	3	23	c = abstr-ana
13	0	32	23	0	76	1	0	15	d = cataphor
3	0	6	1	4	47	0	0	5	e = indiv-vague
18	2	68	26	8	218	2	0	28	f = indiv
0	0	0	2	0	3	2	0	2	g = deictic
0	0	3	0	0	4	0	0	0	h = textual-deictic
6	0	44	9	3	77	1	0	44	i = abandoned

Fig. 4. Confusion matrix for LANCHART+TV dialogues

cially in texts, but they did not have any problem in distinguishing expletives from anaphoric uses of the personal pronouns. Classification also confused a number of individual and abstract anaphora in the texts. This was in few cases also a problematic issue for humans because of the ambiguity of the data. Vague anaphors were often not recognised as such, but this is understandable because they often occur in the same contexts as non-vague anaphors. Finally must classes were mixed up in the LANCHART+TV data.

In table 4 the results obtained for each category by the best performing algorithms on the four datasets are given.
The results of all the experiments indicate that the classification algorithms give significantly better results than the baseline, although the results obtained on multi-party dialogues were much worse than those obtained on the other data. The results with respect to the baseline for the texts, the monologues and the DANPASS dialogues show an improvement of 36.4%, 30.7% and 33.7%, respectively, with respect to the baseline, while the improvement for the LANCHART+TV dialogues is only 19.1%.

Although these results cannot be directly compared with the results reported for the classification of the functions of the Dutch *het* in [12], the magnitude of the improvement with respect to the baseline in the two experiments is similar, except for the results obtained on the LANCHART+TV dialogues which are not as good as the other results. Considering the fact that we look at more categories and more types of data than it was the case in the Dutch experiments, the results we have obtained are positive.

The reasons for the bad results obtained on the LANCHART+TV dialogues compared with the results obtained for the DANPASS data are many. The most important are, in our opinion, the following. Firstly these dialogues are free-conversational and include four discourse participants, while the DANPASS dialogues are two-party MAPTASK dialogues which are much more homogeneous. Secondly the quality of the transcription of the DANPASS dialogues is much higher than that of the transcription of the LANCHART dialogues. In the latter transcriptions there were a number of errors which we did not correct, and the timestamps in the speakers' tracks were not always precisely marked. Because we used these timestamps to automatically determine the order in which simultaneous speech had to be represented in the format required by PALINKA, there are probably a number of errors in the data. Finally, the distribution of the pronominal function types in the LANCHART dialogues is different from that in the other datasets, and the automatic treatment of multi-party dialogues should include information of various type such as the physical objects in the space where the conversation take place, including the discourse participants and adjacency pairs. This type of information was not available for the LANCHART corpus.

The F-measure for the recognition of expletives on the basis of the annotation of the pronominal function is 78.8% in the texts, 30% in the DANPASS monologues, 39% in the DANPASS dialogues and, finally, 32.9% in the LANCHART+TV. Only the measures obtained for the texts are satisfactorily and

function	Precision	Recall	F-measure
NBTree on Texts			
expletive	69.1	91.6	78.8
cataphor	50	19	27.5
indivindual anaphor	61.4	44.6	51:7
abstract anaphor	61	59.2	60.1
vague abstract anaphor	0	0	0
vague individual anaphor	60	37.5	46.2
deictic	0	0	0
textual deictic	0	0	0
SMO on Monologues			
expletive	35.3	26.1	30
cataphor	35.3	24	28.6
individual anaphor	71.9	86.2	78.4
abstract anaphor	81.3	70.3	75.4
vague abstract anaphor	77.8	50	60.9
vague individual anaphor	41.7	22.7	29.4
deictic	0	0	0
abandoned	58.3	56.8	57.5
SMO on DanPASS dialogues			
expletive	42.1	36.4	39
cataphor	27.8	12.8	17.5
individual anaphor	58.1	73.2	64.8
abstract anaphor	68.6	68.2	68.4
vague abstract anaphor	0	0	0
vague individual anaphor	23.3	14.9	18.2
deictic	33.3	14.3	20
textual deictic	0	0	0
abandoned	56	49.5	52.6
KStar on lanchart dialogues			
expletive	30.4	16.9	21.8
cataphor	23.5	14.4	17.8
individual anaphor	35.9	58.9	44.6
abstract anaphor	41.6	42.5	42
vague abstract anaphor	16.7	4.3	6.9
vague individual anaphor	17.4	6.1	9
deictic	28.6	22.2	25
textual deictic	0	0	0
abandoned	35.5	23.9	28.6

Table 4. Classification results per category

near to those obtained in [1] where a lot of features and two word lists were used for identifying non-referential from referential uses of *it*.

In the light of the obtained classification results, we are now revising some of the annotations of the function of pronouns. This is especially the case for the cataphoric function.

4.3 Classification of pronouns in texts enriched with PoS and lemma information

In these experiments we run classification on the texts adding to the words lemma and PoS information. A window of one word preceding and three words following the sn-pronouns was used in order to reduce the size of the data.

The best results obtained by various classifiers on n-grams of words, of words+lemma, of words+PoS and of words+lemma+PoS are in table 5. These

Data	Algorithm	Precision	Recall	F-measure
All	Baseline	18.3	42.8	25.7
word	Rotation Forest	60.7	63.3	60.5
word+lemma	NBTree	61.4	63.9	62
word+PoS	RotationF	62.4	64	61.5
word+lemma+PoS	SMO	61.3	64.3	62.1

Table 5. Classification results: words and linguistic features

results indicate that adding lemma and PoS information increases the performance of classification, but these improvements are not significant[6].

The precision of the PoS tagger (the Brill tagger [2] trained on the Danish Parole corpus [11]) used to tag the textual data is approx. 97%. The precision of the CST lemmatiser [13] which was used on the texts is also approx. 97%.

Using manually corrected annotation might improve more the classification results.

5 Conclusions and future work

In the paper we have described unsupervised and supervised machine learning experiments with the purpose of recognising the function of Danish sn-pronouns in texts and spoken data of various type.

The results of our clustering experiments indicate that unsupervised learning on datasets of the size we are working with does not provide satisfactory results for the task of recognising so fine-grained functions of sn-pronouns as those provided in the annotation because too few clusters are identified and correctness is too low.

[6] In the experiments significance was calculated as corrected resampled t-test via the WEKA experimenter[26].

The results of classification using simple n-grams and the annotation of the function of sn-pronouns gave an improvement with respect to the baseline of 36.4% on text data, 37.9% on the DanPASS monologues and 43.1% on the DanPASS dialogues and 19.1% on the lanchart+tv dialogues. Our results for the first three datasets are better than those reported for a Dutch sn-pronoun by [12]. These results indicate that classifiers can be useful to tag the function of pronouns in texts, monologues and some types of dialogues, although the data cannot be used without manual correction.

We also run the classification experiments on the text data adding lemma and PoS information to the n-grams . The added linguistic information improved the performance of the classifiers on the data, but the improvement is not significant.

An analysis of the human classification of the function of pronouns in light of the results of classification indicates that the definition of the cataphoric function is problematic, and that vague anaphors are in many cases difficult to identify automatically. We are now revising some of the annotations in the light of the classification results.

In future we will include in the data syntactic information extracted from a large computational lexicon which contains some of the information which is useful to identify expletive, abstract and individual anaphoric uses of pronouns and test whether classification improves on our datasets enriched with this type of information.

References

1. Boyd, A., Gegg-Harrison, W., Byron, D.: 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*, pages 40–47, Ann Arbor Michigan, June. (2005)
2. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing. A Case Study in Part of Speech Tagging. *Computational Linguistics*, 21(4):543–565 (1995)
3. Byron, D.K.: Resolving pronominal reference to abstract entities. In *Proceedings of the 40^{th} Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 80–87 (2002)
4. Carletta, J.: Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2):249–254 (1996)
5. Cohen, J.: Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bullettin*, 70:213–220 (1968)
6. Daelemans, W., Hoste, V., De Meulder, F., Naudts, B.: 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)*, pages 84–95, Cavtat-Dubrovnik, Croatia (2003)
7. Eckert, M., Strube, M.: Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89. (2001)
8. Evans, R.: A comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal It. In *NLP 2000*, LNCS 1835, pages 233–240. Springer-Verlag, Berlin Heidelberg. (2000)

9. Gregersen, F.: The LANCHART Corpus of Spoken Danish, Report from a corpus in progress. In *Current Trends in Research on Spoken Language in the Nordic Countries*, pages 130–143. Oulu University Press. (2007)
10. Grønnum, N.: DanPASS - A Danish Phonetically Annotated Spontaneous Speech Corpus. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, Genova, Italy, May. (2006)
11. Hansen, D. H.: *Træning og brug af Brill-taggeren på danske tekster*. Ontoquery technical report, Center for Sprogteknologi, Copenhagen (2000)
12. Hoste, V., Hendrickx, I., Daelemans, W.:Disambiguation of the Neuter Pronoun and Its Effect on Pronominal Coreference Resolution. In V. Matousek and P. Mautner (eds.) *Text, Speech and Dialogue, Proceedings of the 10th International Conference, (TSD 2007)*, volume 4629 of *Lecture Notes in Computer Science*, pages 48–55, Pilsen, Czech Republic, September. Springer.
13. Jongejan, B., Hansen, D. H.: *The CST Lemmatiser* Technical report, Centre for Language Technology (2001)
14. Keson, B., Norling-Christensen, O.: PAROLE-DK. Technical report, Det Danske Sprog- og Litteraturselskab, http://korpus.dsl.dk/e-resurser/parole-korpus.php. (1998)
15. Lyons, J.: *Semantics*, volume I-II. Cambridge University Press. (1977)
16. Maegaard, B., Offersgaard, L., Henriksen, L., Jansen, H., Lepetit, X., Navarretta, C., Povlsen, C.: The MULINCO corpus and corpus platform. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, editors: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pages 2148–2153, Genova, Italy, May. (2006)
17. Mitkov, R., Hallett, C.: Comparing Pronoun Resolution Algorithms. *Computational Intelligence*, 23(2):262–297. (2007)
18. Mitkov, R., Evans, R., Orasan, C.: A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method. In A. Gelbukh, editor, *CICLing 2002*, LNCS 2276, pages 168–186. Springer-Verlag. (2002)
19. Müller, C: Resolving it, this and that in unrestricted multi-party dialog. In *Proceedings of ACL-2007*, pages 816–823, Prague. (2007)
20. Navarretta, C.: *The use and resolution of Intersentential Pronominal Anaphora in Danish Discourse*. Ph.D. thesis, University of Copenhagen, February (2002)
21. Navarretta, C.: Resolving individual and abstract anaphora in texts and dialogues. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 233–239, Geneva, Switzerland (2004)
22. Navarretta, C., Olsen, S.: Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-2008*, Marrakesh, Morocco, May. ELRA (2008).
23. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 730–736, Taipei, Taiwan, August. (2002)
24. Orasan, C.: PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39–43, Sapporo. (2003)
25. Strube, M., Müller, C.: A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the ACL'03*, pages 168–175 (2003)
26. Witten, I.H.; Frank, E.: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition. (2005)