

The Referential versus Non-referential Use of the Neuter Pronoun in Dutch and English

Veronique Hoste¹, Iris Hendrickx² and Lieve Macken¹

⁽¹⁾ LT3, Departement of Translation Studies

Ghent University Association

veronique.hoste@hogent.be

lieve.macken@hogent.be

⁽²⁾ CNTS

University of Antwerp

iris.hendrickx@ua.ac.be

Abstract

This paper discusses a corpus-based investigation of the distribution of the third-person neuter singular pronoun in Dutch (“het”). We labeled all pronominal occurrences of “het” in a large corpus of documents. On the basis of the annotated corpora, we developed an automatic classification system using machine learning techniques to distinguish between the different uses of the neuter pronoun. Although our annotation reveals a completely different distribution of the different uses of the pronoun in Dutch and English, we show that the learning method used for English can be successfully ported to Dutch.

1. Introduction

In computational linguistics, the automatic detection of the expletive versus anaphoric use of the third person neuter pronoun finds its motivation in the task of automatic coreference resolution. It is a research area which is becoming increasingly popular in natural language processing (NLP) research and it is a key task in applications such as machine translation, automatic summarization and information extraction for which text understanding is of crucial importance. When people communicate, they aim for cohesion. Text is therefore “not just a string of sentences. It is not simply a large grammatical unit, something of the same kind as a sentence, but differing from it in size - a sort of supersentence, a semantic unit.” (Halliday and Hasan 1976, p. 291). Coreference, in which the interpretation of an element in conversation depends on a previously mentioned element, is one possible technique to achieve this cohesion, a technique to construct that supersentence. Through the use of shorter or alternative linguistic structures which refer to previously mentioned elements in spoken or written text, coherent communication can be achieved. A good text understanding largely depends on the correct resolution of these coreferential relations.

The goal of this paper is twofold. In a first step, we investigate whether the distribution of the different uses of the Dutch neuter pronoun is similar to the one reported for English. Furthermore, we investigate whether the machine learning approach, successfully applied for the English “it”, can be easily ported to the automatic classification of the Dutch “het”. In order to classify the third person singular neuter pronoun, two different types of approaches have been proposed for English: rule-based strategies (Paice and Husk, 1987; Lappin and Leass, 1994) and

machine learning approaches (Boyd, Gegg-Harrisson and Byron, 2005; Clemente et al., 2004; Evans, 2001; Müller, 2006), which are mostly focused on the distinction between the referential versus non-referential use of the neuter pronoun.

The remainder of this paper is organised as follows. Section 2 introduces the data used for the experiments and describes the annotation and the distribution of phenomena of interest in the data sets. The experimental set-up and results are described in Sections 3 and 4. Section 5 summarizes the main findings of the paper.

2. KNACK-2002 and SPECTRUM

In order to study and automatically model the different uses of the third-person neuter singular pronoun in Dutch, we annotated the following data sets. We labeled all occurrences of “het” in a large corpus of documents (ca. 250,000 tokens) consisting of news magazine texts and medical encyclopedia texts. In order to measure inter-annotator agreement, this labeling was done by two annotators.

Two corpora were annotated with information on the third person neuter pronoun: KNACK-2002, a corpus of news magazine texts (106,011 tokens) and SPECTRUM, a corpus with medical encyclopedia texts (133,887 tokens). The first corpus is based on KNACK, a Flemish weekly news magazine with articles on national and international current affairs. It covers a wide variety of topics in economical, political, scientific, cultural and social news. For the construction of the corpus, we used a selection of articles of different lengths, which all appeared in the first ten weeks of 2002. For a more detailed description of the KNACK-2002 corpus and its annotation, we refer to (Hoste, 2005). The second Dutch corpus is part of the Spectrum medical encyclopedia, namely the first 1000 documents. This data set was licensed to the IMIX project by Spectrum, but can also be viewed online at: <http://www.kiesbeter.nl/medischeinformatie/Page/MedischWoordenboek.aspx>

Two linguists annotated the corpora in parallel in accordance with the annotation guidelines described below, which are based on the general Dutch grammar (ANS)¹. As input, the annotators received free text in which all occurrences of “het” were marked, the majority of which involved “het” as definite article. For the annotation of the personal pronoun “het”, the annotators had to differentiate between the non-referential use of the pronoun (NOREF) as in example (1) and its referential use. In example (1), “het” is part of an idiomatic expression and not referential.

- (1) Leopold haalt scherp uit naar onder meer Hubert Pierlot, de eeuwige zondebok met wie hij **het** niet kon vinden.
English: Leopold sharply attacks among others Pierlot, the eternal scapegoat with whom he can't get on.

We distinguished between the following four types of referential use: (i) reference to preceding “het” words (REF_NP) as in example (2), (ii) reference to a preceding clause (REF_SENT) as in example (3), (iii) “het” as anticipatory subject (REF_LOOS_SUBJ) as in (4) or as anticipatory object (REF_LOOS_OBJ) and

¹ <http://oase.uci.ru.nl/~ans/>

finally, “het” as subject of a nominal predicate (REF_PRED) as in (5). The referential “het” is marked in bold, whereas the antecedent is printed in italics.

- (2) Pakistan is *een samengesteld land*, **het** bij elkaar houden is altijd een prioriteit geweest.
English: Pakistan is a compound country; keeping it together has always been a priority.
- (3) Op een soorgelijke manier *zijn andere stromingen en stijlrichtingen verweven tot zinvolle verbanden*. Dat **het** niet altijd even uitgesponnen of intens gebeurt, komt omdat er gewoon minder grote groepen werken van voorradig zijn.
English: In a similar manner, other movements and schools have been tied up to significant connections. The fact that it has not always happened equally intensive, is due to the less large groups of available works.
- (4) Om al deze redenen, en om het afglijden van een belangrijk en Centraal-Afrikaans land naar de dictatuur af te remmen, leek **het** aangewezen *om naar de aanstaande verkiezingen zoveel mogelijk internationale waarnemers te sturen, en zoveel mogelijk pers*.
English: For all these reasons, and to prevent an important Central African country from slipping further and further into a dictatorship, it seemed appropriate to send to the forthcoming elections as many international observers and press as possible.
- (5) “Wat een mooie hond, mevrouw. Van welk ras is hij?” Waarop mijn moeder antwoordde: “**Het** is *geen hond*, meneer de eerste minister, **het** is *een leeuw*.”
English: “What a beautiful dog, madam. What breed is it?” To which my mother replied: “It is not a dog, Mr. Prime Minister, it’s a lion.”

On the Knack-2002 data, a kappa agreement score was obtained of 0.74. Table 1 gives the contingency table for the Knack-2002 data. The diagonal cells, which represent the agreements between the two annotators, show that the pronominal “het” is predominantly used as a non-referential pronoun. The off-diagonal cells show that for the Knack data the two annotators mainly disagree on the ref_pred versus ref_loos_subj use and on the ref-sent versus non-referential use. On the Spectrum data, the kappa score was 0.81.

	<i>ART</i>	<i>NOREF</i>	<i>REF_ LOOS_ OBJ</i>	<i>REF_ LOOS_ SUBJ</i>	<i>REF_ NP</i>	<i>REF_ PRED</i>	<i>REF_ SENT</i>	<i>NOTYPE</i>	
<i>ART</i>	2340	35	5	13	9	4	1	20	2427
<i>NOREF</i>	9	131	5	5	4	2	11	2	169
<i>REF_ LOOS_ OBJ</i>	0	5	12	0	1	0	1	0	19
<i>REF_ LOOS_ SUBJ</i>	2	6	0	60	1	1	0	1	71
<i>REF_ NP</i>	2	5	1	1	80	2	1	1	93
<i>REF_ PRED</i>	5	8	0	19	3	19	2	1	57
<i>REF_ SENT</i>	4	15	2	2	3	3	22	2	53
<i>NOTYPE</i>	2	0	0	0	1	0	0	0	3
	2364	205	25	100	102	31	38	27	2892

Table 1: Contingency table reflecting the inter-annotator agreement on the Knack-2002 data

After this first annotation round, both annotators re-annotated the texts jointly in order to reach a consensus annotation. In total, over 6500 occurrences of “het” were annotated, of which 844 are pronominal. Table 2 gives an overview of the distribution of the different uses of the neuter pronoun in both annotated Dutch corpora. The figures show that for the news magazine texts, the pronoun refers to a preceding noun phrase in 20% of the cases, whereas for the medical texts nearly half of the “het” occurrences refer to an NP. Taking the Dutch corpora as a whole, three categories show a similar distribution: the non-referential use (30.1%), the reference to a preceding noun phrase (32.6%) and the neuter pronoun as anticipatory subject (23.3%).

	<i>Knack-2002</i>	<i>Spectrum</i>	<i>Total</i>
Pronominal use	507/2892	337/3670	884/6562
Non-referential	39.0%	17.7%	30.1%
Ref - preceding clause	5.7%	0.3%	3.5%
Ref – noun phrase	21.3%	49.5%	32.6%
Ref. – anticipatory subject	19.9%	28.5%	23.3%
Ref. – anticipatory object	5.1%	1.2%	3.5%
Ref. – nominal predicate	8.9%	3.9%	6.9%

Table 2: Distribution of the pronominal “het” in the Dutch data sets.

In Table 3 we compare the distributional results for the noun phrase reference cases from Table 2 with the distributional information in corpora designed for English, as for example the MUC-6 and MUC-7 corpora and the corpora described by Evans (2001) or Boyd, Gegg-Harrisson and Byron (2005). The table reveals that the English corpora which have been previously used for the automatic classification of “it” all show a large number (>67%) of occurrences of “it” in which the pronoun refers to a preceding noun phrase.

		<i>Ref – noun phrase</i>
Dutch	Knack- 2002	21.3%
	Spectrum	49.5%
English	MUC-6	74.4%
	MUC-7	80.7%
	Evans(2001)	67.9%
	Boyd, Gegg-Harrison and Byron (2005)	69.9%

Table 3 : Number of times “het”/”it” refer to a preceding noun phrase

3. Experimental Setup

3.1. Preprocessing

In order to classify the third person singular neuter pronoun, two different types of approaches have been proposed earlier (all for English): rule-based strategies as proposed by Paice and Husk (1987) and Lappin and Leass (1994) and machine learning approaches as in Boyd, Gegg-Harrison and Byron (2005) or Evans (2001).

For the construction of the machine learning data sets, the following preprocessing steps were taken. Tokenization was performed by a rule-based system using regular expressions. Lemmatization was performed using a memory-based lemmatizer trained on a lexicon derived from the Spoken Dutch Corpus (CGN), a 10-million word corpus of spoken Dutch². Part-of-speech tagging and text chunking were performed by the memory-based tagger MBT (Daelemans and van den Bosch, 2005), which was also trained on the CGN corpus. The part-of-speech classes of the CGN are rich. Apart from defining that a word is a pronoun (VNW), a verb (WW) or something else, a part-of-speech tag contains several other features of the word as illustrated by the preprocessed text in the example below.

Belgische	Belgisch	ADJ(prenom,basis,met-e,stan)	B-NP
militairen	militair	N(soort,mv,basis)	I-NP
weten	weten	WW(pv,tgw,mv)	B-VP
niet	niet	BW()	B-ADVP
waar	waar	VNW(vb,adv-pron,obl,vol,3o,getal)	B-NP
ze	ze	VNW(pers,pron,stan,red,3,mv)	B-NP
zullen	zullen	WW(pv,tgw,mv)	B-VP
worden	worden	WW(adv,vrij,zonder)	B-VP
ingezet	inzetten	WW(vd,vrij,zonder)	B-VP
en	en	VG(neven)	B-VG
dus	dus	BW()	B-ADVP
is	zijn	WW(pv,tgw,ev)	B-VP
het	het	VNW(pers,pron,stan,red,3,ev,onz)	B-NP
van	van	VZ(init)	B-PP
belang	belang	N(soort,ev,basis,onz,stan)	B-NP

² <http://lands.let.ru.nl/cgn>

dat	dat	VG(onder)	B-VG
er	er	VNW(aanw,adv-pron,stan,red,3,getal)	B-NP
in	in	VZ(init)	B-PP
potentiële	potentieel	ADJ(prenom,basis,met-e,stan)	B-NP
conflictzones	conflictzone	N(soort,mv,basis)	I-NP
contacten	contact	N(soort,mv,basis)	I-NP
worden	worden	WW(pv,tgw,mv)	B-VP
gelegd	leggen	WW(vd,vrij,zonder)	B-VP
.	.	LET()	O

The information obtained through this preprocessing was used in the construction of the feature vectors for our learning techniques. These feature vectors consist of attribute/value pairs which contain possibly disambiguating information for the classifier, whose task it is to accurately predict the class of novel instances. An ideal feature vector consists of features which are all highly informative and which can lead the classifier to optimal performance.

For all occurrences of “het”, a feature vector was built consisting of 38 features, as shown in the Dutch example below.

1 1 3 9 het no LID(bep,stan,evon) B-NP aangeboren , vaker echter is
aangeboren , vaak echter zijn ADJ(vrij,basis,zonder) LET()
ADJ(vrij,comp,zonder) BW() WW(pv,tgw,ev) het gevolg van een staaroperatie
het gevolg van een staaroperatie LID(bep,stan,evon) N(soort,ev,basis,onz,stan)
VZ(init) LID(onbep,stan,agr) N(soort,ev,basis,zijd,stan) no REF_NP

These features include positional information (sentence number and position in sentence), information on the focus word itself (word form, part-of-speech and chunk information), furthermore information on the word form, lemma and part-of-speech of five words before and after the focus word, and finally information on the use of a preposition before the focus word (Paice and Husk, 1987). Based on the assumption that verbs which occur more often with “het” indicate the non-anaphoric use of the pronoun, we included for Dutch a last feature for which the association strength was calculated between “het” as a subject and its accompanying verb. This association strength was represented by mutual information scores and was based on the Dutch Twente News Corpus (500 million words). A minimal cut-off frequency of 1000 was chosen.

3.2. Memory-based learning

For the classification of the different uses of the neuter pronoun, we used a memory-based learning algorithm, as was also previously applied to this task by Evans (2001) and Boyd, Gegg-Harrison and Byron (2005).

A memory-based learning (MBL) system consists of two components: a memory-based learning component and a similarity-based performance component. During learning, the learning component adds new training instances to the memory without

any abstraction or restructuring (“lazy learning”). At classification time, the algorithm classifies new instances by searching for the nearest neighbors to the new instance using a similarity metric, and extrapolating from their class.

In our experiments we use the TIMBL (Daelemans and van den Bosch, 2005) software package³ that implements a version of the k nearest neighbour algorithm optimised for working with linguistic datasets and that provides several similarity metrics and variations of the basic algorithm. Although the package provides sensible default settings, which have been validated on a number of data sets, it is by no means certain that they are also the optimal settings for our specific task. Furthermore, although we selected features which we believe to be helpful in disambiguating between the different uses of “het”, it is by no means certain that these features are equally informative. Therefore, we performed joint feature selection and parameter optimization by means of a generational genetic algorithm as described in Hoste (2005).

Joint feature selection and parameter optimization involves searching the space of all possible feature subsets and parameter settings to identify the combination that is optimal or near-optimal. Since exhaustive search is computationally hard for large data sets, genetic algorithms can be used to search large search spaces. Genetic algorithms are search methods based on the mechanics of natural selection and genetics. They require two things: Darwinian fitness-based selection and diversity. The principle behind GAs is quite simple: search starts from a population of individuals which all represent a solution to the problem to be solved. Applied to our problem of the classification of the different uses of “het”/“it”, the problem to be solved will be joint feature selection and parameter optimization. These individuals are typically represented as bit strings of fixed length, called a “chromosome” or “genome”.

Given the modest size of the data sets, all experiments were performed using leave-one-out cross-validation. This implies that the data are divided in as many subsets as there are instances. The algorithm is trained on the concatenation of all subsets minus one. The performance of the trained learning method is tested on the omitted instance. This loop is performed as many times as there are instances. The final performance of the learning method is the average over all performed tests. For a detailed description of the feature selection and parameter optimization experiments, we refer to Hoste, Hendrickx and Daelemans (2007). The results reported in Section 4 are the classification results after optimization.

4. Experimental results

Table 4 gives an overview of the overall and 5-ary classification results of the optimized memory-based classifier. As baseline score, the most frequent class, i.e. reference to a preceding NP, was taken and kept constant over all data sets.

Performance is reported in terms of accuracy, precision, recall and F-score of TIMBL on the three data sets. The accuracy scores measure the overall performance of the classifier. It is the number of correct classifications given by the system divided by the total number of test instances. We also evaluated the performance of the learner on the

³ <http://ilk.uvt.nl>

different uses of “het”. These results are represented in terms of precision, recall and F-score (van Rijsbergen, 1979).

- *Recall* is a measure of the coverage of the system and is obtained as follows:

$$R = \text{number of correct answers given by the system} / \text{total number of possible correct answers in the data set.}$$
- *Precision*, on the other hand, is a measure of how much of the information that is given by the system is actually correct.

$$P = \text{number of correct answers given by the system} / \text{number of answers given by the system.}$$
- *F-measure*, is a combined measure which balances precision and recall by using a parameter β . In our experiments, the β parameter was set to one, which implies that precision and recall receive an equal weight.

$$F = (\beta^2 + 1) * \text{precision} * \text{recall} / \beta^2 * (\text{precision} + \text{recall})$$

Table 4 shows a 30% improvement over the most frequent sense accuracy for the three data sets. The results also show that for some subtypes of referential use, there is too little evidence in the training data to train an accurate classifier on. The non-referential use of “het”, on the other hand, can be detected with a reliability of >70%. For the “het” which refers to a preceding noun phrase, divergent F-scores are obtained: 39.1% for Knack-2002, as opposed to 83.4% for Spectrum.

	<i>Knack-2002</i>			<i>Spectrum</i>			<i>Total</i>		
Baseline	22.3			49.5			32.6		
Accuracy	57.4			78.3			64.8		
	P	R	F	P	R	F	P	R	F
Non-referential	60.2	89.4	71.9	85.1	71.4	77.7	67.5	75.2	71.1
Ref - preceding clause	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Ref – noun phrase	49.3	32.4	39.1	78.7	88.6	83.4	63.2	69.8	66.3
Ref. – anticipatory subject	52.3	55.4	53.8	74.5	79.2	76.8	62.7	73.6	67.8
Ref. – anticipatory object	100	42.3	59.5	0.0	0.0	0.0	100	23.3	37.8
Ref. – nominal predicate	54.5	26.7	35.8	0.0	0.0	0.0	66.7	20.7	31.6

Table 4: Performance in terms of accuracy, precision, recall and F-score of TIMBL on the three Dutch data sets. Both the overall and 5-ary classification results of the optimized memory-based classifier are given.

5. Concluding remarks

In this paper we described the results of an annotation experiment in which the different uses of the third person singular neuter pronoun were marked in two Dutch corpora. Since this work is motivated by the task of pronominal coreference resolution, we focused on the annotation of the neuter pronoun referring to a previous noun phrase. We showed large distributional differences between the same phenomenon in Dutch and English.

We developed a machine learning based system for the disambiguation of referential or non-referential use of “het” using memory-based learning and genetic algorithm based joint optimization of feature selection and algorithm parameter selection. These experiments show the portability of this approach which have been successfully applied to English earlier.

References

Boyd, A., W. Gegg-Harrison and D. Byron (2005) 'Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns'. *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, 40-47.

Baayen, R. H., R. Piepenbrock, H. van Rijn. (1993) 'The CELEX lexical data base on CD-ROM'. Linguistic Data Consortium, Philadelphia, PA.

Clemente, J. C., K. Torisawa and K. Satou (2004) 'Improving the identification of non-anaphoric it using support vector machines'. *Proceedings of the Intl. Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.

Daelemans, W. and A. van den Bosch (2005) *Memory-based Language Processing*. Cambridge: Cambridge University Press

Daelemans, W., J. Zavrel, P. Berck, and S. Gillis (1996) 'MBT: A memory-based part of speech tagger generator'. *Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora*, 14-27.

Evans, R (2001) 'Applying machine learning toward an automatic classification of it'. *Literary and Linguistic Computing*, 45-57.

Halliday, M. and R. Hasan (1976) *Cohesion in English*. London: Longman.

Hirschman, L. and N. Chinchor (1998) Muc-7 coreference task definition. version 3.0'. *Proceedings of the Seventh Message Understanding Conference(MUC-7)*.

Hoste, V. (2005) *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, Antwerp University

Hoste, V., I. Hendrickx and W. Daelemans (2007) Disambiguation of the neuter pronoun and its effect on pronominal coreference resolution. *Text, Speech and Dialogue. Proceedings of the 10th International Conference TSD 2007*, Plzen, Czech Republic, 2007 (LNCS). To appear.

Lappin, S. and H. Leass (1994) 'An algorithm for pronominal anaphora resolution'. *Computational Linguistics*, 535-561.

Müller, C. (2006) 'Automatic detection of nonreferential it in spoken multi-party dialog.' *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*.

'MUC-6. Coreference task definition. version 2.3' (1995) *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pp 335-344.

Paice, C. and G. Husk (1987) 'Towards an automatic recognition of anaphoric features in English text: the impersonal pronoun 'it''. *Computer Speech and Language*, 109-132.

Van den Bosch, A, W. Daelemans and A. Weijters (1996) 'Morphological analysis as classification: an inductive-learning approach'. *Proceedings of the Second International Conference on New Methods in Natural Language Processing*. 79-89.

van Rijsbergen, C.J. (1979) *Information Retrieval*. Butterworth, London.