Chapter 7
COREA: Coreference Resolution for Extracting Answers for Dutch

Iris Hendrickx, Gosse Bouma, Walter Daelemans, and Véronique Hoste

7.1 Introduction

Coreference resolution is essential for the automatic interpretation of text. It has been studied mainly from a linguistic perspective, with an emphasis on the recognition of potential antecedents for pronouns. Many practical NLP applications such as information extraction (IE) and question answering (QA), require accurate identification of coreference relations between noun phrases in general. In this chapter we report on the development and evaluation of an automatic system for the robust resolution of referential relations in text. Computational systems for assigning such relations automatically, require the availability of a sufficient amount of annotated data for training and testing. Therefore, we annotated a Dutch corpus of 100K words with coreferential relations, and in addition we developed guidelines for the manual annotation of coreference relations in Dutch.

I. Hendrickx (✉)
Centro de Linguística da Universidade de Lisboa, Faculdade das Letras, Av. Prof. Gama Pinto 2, 1649-003, Lisboa, Portugal
e-mail: iris@clul.ul.pt

G. Bouma
Information Science, University of Groningen, Oude Kijk in 't Jatstraat 26, Postbus 716, NL 9700 AS, Groningen, The Netherlands
e-mail: g.bouma@rug.nl

W. Daelemans
CLiPS, University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium
e-mail: walter.daelemans@ua.ac.be

V. Hoste
LT3, School of Translation Studies, University College Ghent and Faculty of Linguistics, Ghent University, Groot-Brittanniëlaan 45, B-9000, Gent, Belgium
e-mail: veronique.hoste@hogent.be

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_7,

We evaluated the automatic coreference resolution module in two ways. On the one hand we used the standard internal approach to evaluate a coreference resolution system by comparing the predictions of the system to a hand-annotated gold standard test set. On the other hand we performed an application-oriented evaluation of our system by testing the usefulness of coreference relation information in an NLP application. We ran experiments with a relation extraction module for the medical domain, and measured the performance of this module with and without the coreference relation information. In a separate experiment we also evaluated the effect of coreference information produced by another simple rule-based coreference module in a question answering application.

The chapter is structured as follows. We first summarise related work in Sect. 7.2. We present the corpus that is manually annotated with coreference relations in Sect. 7.3.1. Section 7.3.2 details the automatic coreference resolution system and in Sect. 7.4 we show the results of both the internal and the application-oriented evaluation. We conclude in Sect. 7.5.

7.2 Related Work

In the last decade considerable efforts have been put in annotating corpora with coreferential relations. For English, many different data sets with annotated coreferential relations are available such as the MUC-6 [8] and MUC-7 [23] data sets, ACE-2 [7], GNOME corpus [26], ARRAU [27], and more recently, OntoNotes 3.0 [39]. But also for other languages data sets exist such as for German, the TBa-D/Z coreference corpus [12] and the Potsdam corpus [19], for Czech the Prague Dependency Treebank (PDT 2.0) [20], for Catalan AnCora-CO [29], for Italian I-CAB [22] and the Live Memories Corpus [31], and the Copenhagen Dependency Treebank [18] for Danish, English, German, Italian, and Spanish. Most of these corpora follow their own annotation scheme. In SemEval-2010, Task 1 *Coreference Resolution in Multiple Languages* was devoted to multi-lingual coreference resolution for the languages Catalan, Dutch, English, German, Italian and Spanish [30]. The CoNLL 2011 and 2012 shared tasks are also dedicated to automatic coreference resolution.

For Dutch, besides the COREA corpus described in Sect. 7.3.1 there is currently also a data set of written new media texts such as blogs [9] developed in the DuOMan project described in Chap. 20, page 359 and a substantial part (one million words) of the SoNaR corpus [32] (Chap. 13, page 219) is also annotated with coreference. All these data sets have been annotated according to the COREA annotation guidelines. For the Dutch language we can now count on a large, and rich data set that is suitable both for more theoretical linguistic studies of referring expressions and for practical development and evaluation of coreference resolution systems. By covering a variety of text genres, the assembled data set can even be considered as a unique resource for cross-genre research.

Currently there are not many coreference resolution systems for Dutch available. The first full-fledged system was presented by Hoste [14, 15] in 2005 and this is

the predecessor of the system described in Sect. 7.3.2. More recently, two of the participating systems in the SemEval-2010 Task 1 on multi-lingual coreference resolution were evaluated for all six languages including Dutch. The UBIU system [40], was a robust language independent system that used a memory-based learning approach using syntactic and string matching features. The SUCRE system [17] obtained the overall best results for this SemEval task and used a more flexible and rich feature construction method and a relational database in combination with machine learning.

7.3 Material and Methods

7.3.1 Corpus and Annotation

The COREA corpus is composed of texts from the following sources:

- Dutch newspaper articles gathered in the DCOI project[1]
- Transcribed spoken language material from the Spoken Dutch Corpus (CGN)[2]
- Lemmas from the Spectrum (Winkler Prins) medical encyclopedia as gathered in the IMIX ROLAQUAD project[3]
- Articles from KNACK [16], a Flemish weekly news magazine.

All material from the first three sources was annotated in the COREA project. The material from KNACK was already annotated with coreference relations in a previous project (cf. [14]). Note that the corpus covers a number of different genres (speech transcripts, news, medical text) and contains both Dutch and Flemish sources. The latter is particularly relevant as the use of pronouns differs between Dutch and Flemish [36].

The size of the various subcorpora, and the number of annotated coreference relations is given in Table 7.1.

For the annotation of coreference relations we developed a set of annotation guidelines [3] largely based on the MUC-6 [8] and MUC-7 [23] annotation scheme for English. Annotation focuses primarily on coreference or IDENTITY relations between noun phrases, where both noun phrases refer to the same extra-linguistic entity. These multiple references to the same entity can be regarded as a *coreferential chain* of references. While these form the majority of coreference relations in our corpus, there are also a number of special cases. A BOUND relation exists between an anaphor and a quantified antecedent, as in *Everybody$_i$ did what they$_i$ could*. A BRIDGE relation is used to annotate part-whole or set-subset relations, as in *the tournament$_i$ ... the quarter finals$_i$*. We also marked predicative (PRED) relations,

[1] DCOI: http://lands.let.ru.nl/projects/d-coi/

[2] CGN: http://lands.let.ru.nl/cgn/

[3] IMIX: http://ilk.uvt.nl/rolaquad/

Table 7.1 Corpus statistics for the coreference corpora developed and used in the COREA project. IDENT, BRIDGE, PRED and BOUND refer to the number of annotated identity, bridging, predicative, and bound variable type coreference relations respectively

Corpus	DCOI	CGN	MedEnc	Knack
#docs	105	264	497	267
#tokens	35,166	33,048	135,828	122,960
# IDENT	2,888	3,334	4,910	9,179
# BRIDGE	310	649	1,772	na
# PRED	180	199	289	na
# BOUND	34	15	19	43

as in *Michiel Beute$_i$ is a writer$_i$*. Strictly speaking, these are not coreference relations, but we annotated them for a practical reason. Such relations express extra information about the referent that can be useful for example for a question answering application. We used several attributes to indicate situations where a coreference relation is in the scope of negation, is modified or time dependent, or refers to a meta-linguistic aspect of the antecedent.

Annotation was done using the MMAX2 tool.[4] For the DCOI and CGN material, manually corrected syntactic dependency structures were available. Following the approach of [12], we used these to simplify the annotation task by creating an initial set of markables beforehand. Labeling was done by several linguists.

To estimate the inter-annotator agreement for this task, 29 documents from CGN and DCOI were annotated independently by two annotators, who marked 517 and 470 coreference relations, respectively. For the IDENT relation, we compute inter-annotator agreement as the F-measure of the MUC-scores [38] obtained by taking one annotation as 'gold standard' and the other as 'system output'. For the other relations, we compute inter-annotator agreement as the average of the percentage of *anaphor-antecedent* relations in the gold standard for which an *anaphor-antecedent′* pair exists in the system output, and where *antecedent* and *antecedent′* belong to the same cluster (w.r.t. the IDENT relation) in the gold standard. Inter-annotator agreement for IDENT is 76 % F-score, for bridging is 33 % and for PRED is 56 %. There was no agreement on the three BOUND relations marked by each annotator. The agreement score for IDENT is comparable, though slightly lower, than those reported for comparable tasks for English and German [13, 37]. Poesio and Vieira [28] reports 59 % agreement on annotating 'associative coreferent' definite noun phrases, a relation comparable to our BRIDGE relation.

The main sources of disagreement were cases where one of the annotators fails to annotate a relation, where there is confusion between PRED or BRIDGE and IDENT, and various omissions in the guidelines (i.e. whether to consider headlines and other leading material in newspaper articles as part of the text to be annotated).

[4] http://mmax2.sourceforge.net/

7.3.2 Automatic Resolution System

We developed an automatic coreference resolution tool for Dutch [14] that follows the pairwise classification method of potential anaphora-antecedent pairs similar to the approach of Soon et al. [33]. As supervised machine learning method we decided to use memory-based learning. We used the Timbl software package (version 5.1) [4] that implements several memory-based learning algorithms.

As we used a supervised machine learning approach to coreference resolution the first step was to train the classifier on examples of the task at hand: texts with manually annotated coreference relations. These manually annotated texts needed to be transformed into training instances for the machine learning classifier. First the raw texts were preprocessed to determine the noun phrases in the text and to gather grammatical, positional, and semantic information about these nouns. This preprocessing step involved a cascade of NLP steps such as tokenisation, part-of-speech tagging, text chunking, named entity recognition and grammatical relation finding as detailed in Sect. 7.3.3.

On the basis of the preprocessed texts, training instances were created. We considered each noun phrase (and pronoun) in the text as a potential anaphor for which we needed to find its antecedent. We processed each text backward, starting with the last noun phrase and pairing it with each preceding noun phrase, with a restriction of 20 sentences backwards. Each pair of two noun phrases was regarded as a training instance for the classifier. If a pair of two noun phrases belonged to the same manually annotated coreferential chain, it got a positive label; all other pairs got a negative label. For each pair a feature vector was created to describe the noun phrases and their relation (detailed in Sect. 7.3.4). Test instances were generated in the same manner. In total, 242 documents from the KNACK material were used as training material for the coreference resolution system.

The output from the machine learning classifier was a set of positively classified instances. Instead of selecting one single antecedent per anaphor (such as for example [25, 33]), we tried to build complete coreference chains for the texts and reconstruct these on the basis of the positive instances. As we paired each noun phrase with every previous noun phrase, multiple pairs can be classified as positive. For example, we have a text about Queen Beatrix and her name is mentioned five times in the text. In the last sentence there is the pronoun "she" referring to Beatrix. So we have a coreferential chain in the text of six elements that all refer to the same entity Beatrix. If we create pairs with this pronoun and all previous noun phrases in the text, we will have five positive instances each encoding the same information: "she" refers to Beatrix. For the last mention of the name Beatrix, there are four previous mentions that also refer to Beatrix, leading to four positive instances. In total there are $5 + 4 + 3 + 2 + 1 = 15$ positive instances for this chain while we need a minimum of five pairs to reconstruct the coreferential chain. Therefore we needed a second step to construct the coreferential chains by grouping and merging the positively classified instances that cover the same noun phrases. We grouped pairs

together and computed their union. When the overlap was larger than 0.1 we merged the chains together (we refer to [10] for more details on the merging procedure).

7.3.3 Preprocessing

The following preprocessing steps were performed on the raw texts: First, tokenisation was performed by a rule-based system using regular expressions. Dutch named entity recognition was performed by looking up the entities in lists of location names, person names, organisation names and other miscellaneous named entities. We applied a part-of-speech tagger and text chunker for Dutch that used the memory-based tagger MBT [5], trained on the Spoken Dutch Corpus.[5] Finally, grammatical relation finding was performed, using a shallow parser to determine the grammatical relation between noun chunks and verbal chunks, e.g. subject, object, etc. The relation finder [34] was trained on the previously mentioned Spoken Dutch Corpus. It offered a fine-grained set of grammatical relations, such as modifiers, verbal complements, heads, direct objects, subjects, predicative complements, indirect objects, reflexive objects, etc. We used the predicted chunk tags to determine the noun phrases in each text, and the information created in the preprocessing phase was coded as feature vectors for the classification step.

7.3.4 Features

For each pair of noun phrases we constructed a feature vector representing their properties and their relation [14]. For each potential anaphor and antecedent we listed their individual lexical and syntactic properties. In particular, for each potential anaphor/antecedent, we encode the following information, mostly in binary features:

- Yes/no pronoun, yes/no reflexive pronoun, type of pronoun (first/second/third person or neutral),
- Yes/no demonstrative,
- Type of noun phrase (definite or indefinite),
- Yes/no proper name,
- Yes/no part of a named entity,
- Yes/no subject, object, etc., of the sentence as predicted by the shallow parser.

For the anaphor we also encoded its local context in the sentence as a window in words and PoS-tags of three words left and right of the anaphor. We represented the relation between the two noun phrases with the following features:

[5] http://lands.let.ru.nl/cgn

- The distance between the antecedent and anaphor, measured in noun phrases and sentences;
- Agreement in number and in gender between both;
- Are both of them proper names, or is one a pronoun and the other a proper name;
- Is there a complete string overlap, a partial overlap, a overlap of the head words or is one an abbreviation of the other.

One particularly interesting feature that we have explored was the usage of semantic clusters [35]. These clusters were extracted with unsupervised k-means clustering on the Twente Nieuws Corpus.[6] The corpus was first preprocessed by the Alpino parser [1] to extract syntactic relations. The top-10,000 lemmatised nouns (including names) were clustered into a 1,000 groups based on the similarity of their syntactic relations. Here are four examples of the generated clusters:

- {barrière belemmering drempel hindernis hobbel horde knelpunt obstakel struikelblok} (Eng: obstacle impediment threshold hindrance encumbrance hurddle knot obstacle)
- {Disney MGM Paramount PolyGram Time_Warner Turner Viacom }
- {Biertje borrel cocktail cola drankje glaasje kopje pilsje} (Eng:beer shot cocktail glass cup cola drink pils)
- {Contour schaduw schim schrikbeeld silhouet verhaallijn} (Eng:contour shade shadow chimera silhouette story-line)

For each pair of referents we constructed three features as follows. For each referent the lemma of the head word was looked up in the list of clusters. The number of the matching cluster, or 0 in case of no match, was used as the feature value. We also constructed two features presenting the cluster number of each referent and a binary feature marking whether the head words of the referents occur in the same cluster or not.

In the first version of the coreference resolution system we coded syntactic information as predicted by the memory-based shallow parser in our feature set of 47 features [14]. In the COREA project we also investigated whether the richer syntactic information of a full parser would be a helpful information source for our task [11]. We used the Alpino parser [1], a broad-coverage dependency parser for Dutch to generate the 11 additional features encoding the following information:

- Named Entity label as produced by the Alpino parser, one for the anaphor and one for the antecedent.
- Number agreement between the anaphor and antecedent, presented as a four valued feature (values: *sg, pl, both, measurable_nouns*).
- Dependency labels as predicted for (the head word of) the anaphor and for the antecedent and whether they share the same dependency label.
- Dependency path between the governing verb and the anaphor, and between the verb and antecedent.

[6] http://www.vf.utwente.nl/$\sim$druid/TwNC/TwNC-main.html

- Clause information stating whether the anaphor or antecedent is part of the main clause or not.
- Root overlap encodes the overlap between 'roots' or lemmas of the anaphor and antecedent. In the Alpino parser, the root of a noun phrase is the form without inflections. Special cases were compounds and names. Compounds are split[7] and we used the last element in the comparison. For names we took the complete strings.

In total, each feature vector consisted of 59 features. In the next section we describe how we selected an optimal feature set for the classification and the results of the automatic coreference resolution experiments with and without these deeper syntactic features.

7.4 Evaluation

We performed both a direct evaluation and an external, application-oriented evaluation. In the direct evaluation we measured the performance of the coreference resolution system on a gold-standard test set annotated manually with coreference information. In the application-oriented evaluation we tried to estimate the usefulness of the automatically predicted coreference relations for NLP applications.

7.4.1 Direct Evaluation

Genetic algorithms (GA) have been proposed [6] as an useful method to find an optimal setting in the enormous search space of possible parameter and feature set combinations. We ran experiments with a generational genetic algorithm for feature set and algorithm parameter selection of Timbl with 30 generations and a population size of 10.

In this experiment we used ten fold cross validation on 242 texts from Knack. The GA was run on the first fold of the ten folds as running the GA is rather time-consuming. The found optimal setting was then used for the other folds as well. We computed a baseline score for the evaluation of the complete coreference chains. The baseline assigned each noun phrase in the test set its most nearby noun phrase as antecedent.

The results are shown in Table 7.2. Timbl scores well above the baseline in terms of F-score but the baseline has a much higher recall. The differences in F-score at the instance level between the model without and with syntactic features, are small,

[7]The Alpino parser uses various heuristics to determine whether words that are not in its dictionary can be analyzed as compounds. The most important heuristic splits a word in two parts where both parts must be in the dictionary, and the split that gives the longest suffix is chosen.

Table 7.2 Micro-averaged F-scores at the instance level and MUC F-scores at the chain level computed in ten fold cross validation experiments. Timbl is run with the settings as selected by the genetic algorithm (GA) without and with the additional Alpino features

Scoring at the instance level			
	Recall	Precision	F-score
TIMBL, GA	44.8	70.5	54.8
TIMBL, GA, with syntax	48.4	64.1	55.1
MUC scoring at the chain level			
	Recall	Precision	F-score
Baseline	81.1	24.0	37.0
TIMBL, GA	36.8	70.2	48.2
TIMBL, GA, with syntax	44.0	61.4	51.3

but when we look at the score computed at the chain level, we see an improvement of 3 % in F-score. Adding the additional features from the Alpino parser improves overall F-score by increasing the recall at the cost of precision.

7.4.2 Application-Oriented Evaluation

Below, we present the results of two studies that illustrate that automatic coreference resolution can have a positive effect on the performance of systems for information extraction and question answering.

7.4.2.1 Coreference Resolution for Information Extraction

To validate the effect of the coreference resolution system in a practical information extraction application, our industrial partner in this project, *Language and Computing NV*, constructed an information extraction module named *Relation Finder* which can predict medical semantic relations. This application was based on a version of the Spectrum medical encyclopedia (MedEnc) developed in the IMIX ROLAQUAD project, in which sentences and noun phrases were annotated with domain specific semantic tags [21]. These semantic tags denote medical concepts or, at the sentence level, express relations between concepts. Example 7.1 shows two sentences from MedEnc annotated with semantic XML tags. Examples of the concept tags are *con_disease, con_person_feature* or *con_treatment*. Examples of the relation tags assigned to sentences are *rel_is_symptom_of* and *rel_treats*.

Example 7.1.

```
<rel_is_symptom_of id="20">
  Bij <con_disease id="2">asfyxie</con_disease> ontstaat een
```

```
    toestand van
    <con_sympton id="7">bewustzijnverlies</con_sympton>
    en <con_disease id="4>shock</con_disease> (nauwelijks
    waarneembare
    <con_person_feature id="8">polsslag</con_person_feature> en
    <con_body_function id="13">ademhaling</con_body_function>).
  </rel_is_symptom_of>
  <rel_treats id="19">
    Veel gevallen van <con_disease id="6">asfyxie</con_disease>
    kunnen door
    <con_treatment id="14">beademing</con_treatment>, of
    door opheffen van de passagestoornis
    (<con_treatment id="15">tracheotomie</con_treatment>)
    weer herstellen.
  </rel_treats>
```

The core of the Relation Finder was a maximum entropy modeling algorithm trained on approximately 2,000 annotated entries of MedEnc. Each entry was a description of a particular item such as a disease or body part in the encyclopedia and contained on average ten sentences. It was tested on two separate test sets of 50 and 500 entries respectively. Our coreference module predicted coreference relations for the noun phrases in the data. We ran two experiments with the Relation Finder. In the first experiment we used the predicted coreference relations as features and the second one we did not use these features. On the small data set we obtained an F-score of 53.03 % without coreference and 53.51 % with coreference information. On the test set with 500 entries we got a slightly better score of 59.15 % F-score without and 59.60 % with coreference information. So for both test sets we observe a modest positive effect for the experiments using the coreference information.

7.4.2.2 Coreference Resolution for Question Answering

The question answering system for Dutch described in [2] used information extraction to extract answers to frequent questions off-line (i.e. the system tried to find all instances of the `capital` relation in the complete text collection off-line, to answer questions of the form *What is the capital of LOCATION?*). Tables with relation tuples were computed automatically for relations such as age of a person, location and date of birth, founder of an organisation, function of a person, number of inhabitants, winner of a prize, etc.

Using manually developed patterns, the precision of extracted relation instances is generally quite high, but coverage tends to be limited. One reason for this is the fact that relation instances are only extracted between entities (i.e. names, dates, and numbers). Sentences of the form *The village has 10,000 inhabitants* do not contain a ⟨*location,number_of_inhabitants*⟩ pair. If we can resolve the antecedent of *the village*, however, we can extract a relation instance.

To evaluate the effect of coreference resolution for this task, [24] extended the information extraction component of the QA system with a simple rule-based

Table 7.3 Number of relation instances, precision, and number of unique instances (facts) extracted using the baseline system, and using coreference resolution

	Instances	Precision	Facts
No coreference resolution	93,497	86 %	64,627
Pronoun resolution	3,915	40 %	3,627
Resolution on definite NPs	47,794	33 %	35,687
Total	145,141	72 %	103,941

coreference resolution system for pronouns. To resolve definite noun phrases, it used an automatically constructed knowledge base containing 1.3M class labels for named entities to resolve definite NPs.

Table 7.3 shows that, after adding coreference resolution, the total number of extracted facts went up with over 50 % (from 93K to 145K). However, the accuracy of the newly added facts was only 40 % for cases involving pronoun resolution and 33 % for cases involving definite NPs.

In spite of the limited accuracy of the newly extracted facts, we noticed that incorporation of the additional facts led to an increase in performance on the questions from the QA@CLEF 2005 test set of 5 % (from 65 to 70 %). We expect that even further improvements are possible by integrating the coreference resolution system described in Sect. 7.3.2.

7.5 Conclusion

Coreference resolution is useful in text mining tasks such as information extraction and question answering. Using coreference resolution, more useful information can be extracted from text, and that has a positive effect on the recall of such systems. However, it is not easy to show the same convincingly in application-oriented evaluations. The reason for this is that the current state-of-the-art in coreference resolution, based on supervised machine learning, is still weak, especially in languages like Dutch for which not a lot of training data is available. More corpora are needed, annotated with coreference relations.

We presented the main outcomes of the STEVIN COREA project, which was aimed at addressing this corpus annotation bottleneck. In this project, we annotated a balanced corpus with coreferential relations, trained a system on it, and carried out both a direct and application-oriented evaluations.

We discussed the corpus, the annotation and the inter-annotator agreement, and described the construction and evaluation of a coreference resolution module trained on this corpus in terms of the preprocessing and the features used.

We evaluated this coreference resolution module in two ways: with standard cross-validation experiments to compare the predictions of the system to a hand-annotated gold standard test set, and a more practically oriented evaluation to

test the usefulness of coreference relation information in information extraction and question answering. In both cases we observed a small but real positive effect of integrating coreference information, despite the relatively low accuracy of current systems. More accurate coreference resolution systems, should increase the magnitude of the positive effect. These systems will need additional semantic and world knowledge features. We showed the positive effect of richer syntactic features as generated by the Alpino parser, and of semantic features by means of the semantic cluster features we tested.

The annotated data, the annotation guidelines, and a web demo version of the coreference resolution system are available to all and are distributed by the Dutch TST HLT Agency.[8]

Acknowledgements We would like to thank Tim Van de Cruys for sharing his data sets of semantic clusters. We thank Anne-Marie Mineur, Geert Kloosterman, and Language & Computing for their collaboration in the COREA project.

References

1. Bouma, G., van Noord, G., Malouf, R.: Alpino: wide-coverage computational analysis of dutch. In: Computational Linguistics in the Netherlands 2000. Selected Papers from the Eleventh CLIN Meeting, Tilburg, The Netherlands (2001)
2. Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedeman, J.: Linguistic knowledge and question answering. Traitement Automatique des Langues **2**(46), 15–39 (2005)
3. Bouma, G., Daelemans, W., Hendrickx, I., Hoste, V., Mineur, A.: The COREA-project, manual for the annotation of coreference in Dutch texts. University Groningen (2007)
4. Daelemans, W., van den Bosch, A.: Memory-Based Language Processing. Cambridge University Press, Cambridge, UK/New York, USA (2005)
5. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: Mbt: a memory-based part of speech tagger generator. In: Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora, pp. 14–27 Santa Cruz, California, USA (1996)
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, Cavtat-Dubrovnik, Croatia, pp. 84–95 (ECML-2003)
7. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, R., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program tasks, data, and evaluation. In: Proceedings of the LREC 2004, Lisbon, Portugal, pp. 837–840 (2004)
8. Grishman, R., Sundheim, B.: Coreference task definition. version 2.3. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, Maryland, USA pp. 335–344 (1995)

[8] www.tst-centrale.org

9. Hendrickx, I., Hoste, V.: Coreference resolution on blogs and commented news. In: Lalitha Devi, S., Branco, A., Mitkov, R. (eds.) Anaphora Processing and Applications. Lecture Notes in Artificial Intelligence, vol. 5847, pp. 43–53. Springer, Berlin/New York (2009)
10. Hendrickx, I., Hoste, V., Daelemans, W.: Evaluating hybrid versus data-driven coreference resolution. In: Anaphora: Analysis, Algorithms and Application, Proceedings of the DAARC 2007, Lagos, Portugal. Lecture Notes in Artificial Intelligence, vol. 4410, pp. 137–150 (2007)
11. Hendrickx, I., Hoste, V., Daelemans, W.: Semantic and Syntactic Features for Anaphora Resolution for Dutch. Lecture Notes in Computer Science, vol. 4919, pp. 351–361. Springer, Berlin (2008)
12. Hinrichs, E., Kübler, S., Naumann, K.: A unified representation for morphological, syntactic, semantic, and referential annotations. In: Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, Ann Arbor, MI (2005)
13. Hirschman, L., Robinson, P., Burger, J., Vilain, M.: Automating coreference: the role of annotated training data. In: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing. Providence, Rhode Island, USA (1997)
14. Hoste, V.: Optimization issues in machine learning of coreference resolution. Ph.D. thesis, Antwerp University, Antwerp, Belgium (2005)
15. Hoste, V., Daelemans, W.: Learning Dutch coreference resolution. In: Proceedings of the Fifteenth Computational Linguistics in The Leiden. The Netherlands (CLIN 2004) (2005)
16. Hoste, V., de Pauw, G.: Knack-2002: a richly annotated corpus of dutch written text. In: Proceedings of LREC 2006, Genoa, Italy, pp. 1432–1437 (2006)
17. Kobdani, H., Schütze, H.: SUCRE: a modular system for coreference resolution. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 92–95. Association for Computational Linguistics, Uppsala, Sweden (2010)
18. Korzen, I., Buch-Kromann, M.: Anaphoric relations in the copenhagen dependency treebanks. In: Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena Proceedings of the DGfS Workshop, Göttingen, Germany pp. 83–98 (2011)
19. Krasavina, O., Chiarcos, C.: PoCoS – Potsdam coreference scheme. In: Proceedings of the Linguistic Annotation Workshop, pp. 156–163. Association for Computational Linguistics, Prague, Czech Republic (2007)
20. Kučová, L., Hajičová, E.: Coreferential relations in the Prague dependency treebank. In: Proceedings of the DAARC 2004, Azores, Portugal, pp. 97–102 (2004)
21. Lendvai, P.: Conceptual taxonomy identification in medical documents. In: Proceedings of the Second International Workshop on Knowledge Discovery and Ontologies, Porto, Portugal pp. 31–38 (2005)
22. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V.B., Sprugnoli, R.: I-CAB: the Italian content annotation bank. In: Proceedings of LREC 2006, Genoa, Italy, pp. 963–968 (2006)
23. MUC-7: Muc-7 coreference task definition. version 3.0. In: Proceedings of the Seventh Message Understanding Conference (MUC-7), Columbia, Maryland, USA (1998)
24. Mur, J.: Increasing the coverage of answer extraction by applying anaphora resolution. In: Fifth Slovenian and First International Language Technologies Conference (IS-LTC '06), Ljubljana, Slovenia (2006)
25. Ng, V., Cardie, C.: Combining sample selection and error-driven pruning for machine learning of coreference rules. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), Philadelphia, PA, USA, pp. 55–62 (2002)
26. Poesio, M.: Discourse annotation and semantic annotation in the GNOME corpus. In: Proceedings of the ACL Workshop on Discourse Annotation, Barcelona, Spain (2004)
27. Poesio, M., Artstein, R.: Anaphoric annotation in the ARRAU corpus. In: Proceedings of the LREC 2008, Marrakech, Morocco, pp. 1170–1174 (2008)
28. Poesio, M., Vieira, R.: A corpus-based investigation of definite description use. Comput. Linguist. **24**(2), 183–216 (1998)
29. Recasens, M., Martí, M.A.: AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. Lang. Res. Eval. **44**(4), 315–345 (2010)

30. Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y.: Semeval-2010 task 1: coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 1–8. Association for Computational Linguistics, Uppsala, Sweden (2010)
31. Rodríguez, K.J., Delogu, F., Versley, Y., Stemle, E.W., Poesio, M.: Anaphoric annotation of wikipedia and blogs in the live memories corpus. In: Proceedings of the LREC 2010. European Language Resources Association (ELRA), Valletta, Malta (2010)
32. Schuurman, I., Hoste, V., Monachesi, P.: Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. In: Proceedings of the LREC 2010. European Language Resources Association (ELRA), Valletta, Malta (2010)
33. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguist. **27**(4), 521–544 (2001)
34. Tjong Kim Sang, E., Daelemans, W., Höthker, A.: Reduction of Dutch sentences for automatic subtitling. In: Computational Linguistics in The Netherlands 2003. Selected Papers from the Fourteenth CLIN Meeting, Antwerp, Belgium pp. 109–123 (2004)
35. Van de Cruys, T.: Semantic clustering in Dutch. In: Proceedings of the Sixteenth Computational Linguistics in the Netherlands (CLIN), Amsterdam, The Netherlands pp. 17–32 (2005)
36. Vandekerckhove, R.: Belgian Dutch versus Netherlandic Dutch: new patterns of divergence? On pronouns of address and diminutives. Multiling. J. Cross-Cult. Interlang. Commun. **24**, 379–397 (2005)
37. Versley, Y.: Disagreement dissected: vagueness as a source of ambiguity in nominal (co-) reference. In: Ambiguity in Anaphora Workshop Proceedings, pp. 83–89. ESSLLI, Malaga, Spain (2006)
38. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), Columbia, Maryland, USA pp. 45–52 (1995)
39. Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., Houston, A.: OntoNotes release 3.0. LDC2009T24. Linguistic Data Consortium (2009)
40. Zhekova, D., Kübler, S.: UBIU: a language-independent system for coreference resolution. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 96–99. Association for Computational Linguistics, Uppsala, Sweden (2010)