Cross-Domain Dutch Coreference Resolution

Orphée De Clercq, Véronique Hoste
LT3, Language and Translation Technology Team
University College Ghent
Groot-Brittannielaan 45
B - 9000 Gent, Belgium
orphee.declercq@hogent.be
veronique.hoste@hogent.be

Iris Hendrickx
Centre of Linguistics
University of Lisbon
Av. Prof. Gama Pinto, 2
1649-003 Lisbon, Portugal
iris@clul.ul.pt

Abstract

This article explores the portability of a coreference resolver across a variety of eight text genres. Besides newspaper text, we also include administrative texts, autocues, texts used for external communication, instructive texts, wikipedia texts, medical texts and unedited new media texts. Three sets of experiments were conducted. First, we investigated each text genre individually, and studied the effect of larger training set sizes and including genre-specific training material. Then, we explored the predictive power of each genre for the other genres conducting cross-domain experiments. In a final step, we investigated whether excluding genres with less predictive power increases overall performance. For all experiments we use an existing Dutch mention-pair resolver and report on our experimental results using four metrics: MUC, B-cubed, CEAF and BLANC. We show that resolving out-of-domain genres works best when enough training data is included. This effect is further intensified by including a small amount of genre-specific text. As far as the cross-domain performance is concerned we see that especially genres of a very specific nature tend to have less generalization power.

1 Introduction

Coreference resolution is the task of automatically recognizing which words or expressions refer to the same discourse entity in a particular text or dialogue.[1] In the last decade considerable efforts have been put in annotating corpora with coreferential relations. Not only a widespread language such as English (e.g. *ACE-2* (Doddington et al., 2004), *ARRAU* (Poesio and Artstein, 2008), *OntoNotes 3.0* (Weischedel et al., 2009)), but also Czech (*PDT 2.0* (Kučová and Hajičová, 2004)), Catalan (*AnCora-Ca* (Recasens and Martí, 2010)) and Italian (*I-CAB* (Magnini et al., 2006))[2] can now rely on substantial resources for coreference research.

One of the challenges in many current NLP tasks is to test their portability across different domains and languages. This portability to other languages was the main objective of the SemEval 2010 Task on Coreference Resolution in Multiple Languages (Recasens et al., 2010). The issue of domain portability was the focus of the ACL 2010 Workshop on Domain Adaptation for NLP (Daumé III et al., 2010).

In this paper we investigate the performance of an existing mention-pair coreference resolver for Dutch (Hoste, 2005; Hendrickx et al., 2008b) across various text genres. More specifically we want to know whether training on out-of-domain data can be done without performance loss. The above-mentioned corpora designed for coreference resolution consist almost exclusively of text from the same genre, i.e. newspaper texts, and as a consequence resulting coreference resolvers are mostly trained on this particular genre. Moreover, when other genres are included, the acquired data are rather scarce: 25K of dialogues in *ARRAU* (Poesio and Artstein, 2008), 23K manuals in *AnATar* (Hammami et al., 2009) or 50K of annotated blogs in *LiveMemories* (Rodríguez et al., 2010). Another related study is the work of Longo and Todirascu (2010). They analyzed a French corpus (50K) consisting of 5 different text genres to develop genre-specific features; in their study

[1] In this article we only discuss nominal coreference, i.e. which coreferential relations exist between noun phrases (common and proper nouns, pronouns).

[2] For a more complete overview we refer to (Recasens, 2010) and (Poesio et al., forthcoming)

Proceedings of Recent Advances in Natural Language Processing, pages 186–193,
Hissar, Bulgaria, 12-14 September 2011.

they use genre-specific features such as average length of the coreferential chain and average distance separating several mentions of the same referent. An exception to this observation of small datasets is the new OntoNotes 4.0 corpus that is used for the CoNLL 2011 Shared Task on unrestricted coreference resolution, as the corpus contains approximately 1 million words from 5 different text genres.[3] We do see a growing interest in one specific different text genre, namely biomedical text in many NLP tasks, including coreference resolution (e.g Yang et al. (2004), Gasperin and Briscoe (2008), Ngan Nguyen and Tsujii (2008)).

The data for the experiments come from three Dutch corpus projects in which coreference was annotated: COREA (Hendrickx et al., 2008a), DuOMAn (Hendrickx and Hoste, 2009) and SoNaR (Schuurman et al., 2010)[4]. Combining these three resources allows us to work with diverse data spread over different text genres. Another advantage is that all data was annotated following the same approach: first all NPs were pre-tagged based on syntactic dependency structures (Bouma and Kloostermans, 2007) and secondly the COREA guidelines (Bouma et al., 2007) were reused in each project. Though the emphasis in this study is on edited text, we also include unedited text, viz. blogs and news comments (Hendrickx and Hoste, 2009). With this cross-domain portability study, we aim to see which genres perform better or worse and whether it is possible to determine a priori which training data to add to our resolver so as to obtain better results. The results are presented using four of the more frequently used evaluation metrics for coreference research, namely MUC (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998), CEAF (Luo and Zitouni, 2005) and BLANC (Recasens and Hovy, 2011).

We show that adding more data to training proves mostly beneficial, especially when genre-specific information is included. Moreover, training a resolver on each genre separately allows us to classify each genre as having good or bad generalization power when applied to other genres. This led us to conduct experiments in which we train on all genres while progressively leaving out the worst-performing cross-domain genres as an attempt to boost overall performance. Although the results are sometimes better, performance does not rise nor drop dramatically. We show that inclusion of some genre-specific training material is necessary, especially when less generalizable genres are to be labeled. However, most effect is perceived by adding more data to training.

The remainder of this paper is organized as follows. In Section 2, we present the datasets and experimental setup of our system and briefly discuss the different evaluation metrics. In Section 3 the results are presented and analyzed, and we report on our experience with the different evaluation metrics. Section 4 concludes this paper by formulating some conclusions and prospects for future work.

2 Datasets and Experimental Setup

In the present study, we aim to investigate the cross-genre portability of an existing mention-pair coreference resolver for Dutch. In order to do so, our system's performance was compared on eight datasets: administrative texts (ADM), autocues (AUTO), texts used for external communication (EXT), instructive texts (INST), journalistic texts (JOUR), medical texts (MED), wikipedia (WIKI), and unedited text (DUO). All data were manually annotated using the COREA guidelines (Bouma et al., 2007). These guidelines allow for the annotation of four relations and special cases are flagged. The four annotated relations are identity (NPs referring to the same discourse entity), bound (expressing properties of general categories), bridge (as in part-whole, superset-subset relations) and predicative. The following special cases were flagged: negations and expressions of modality, time-dependency and identity of sense (as in the so-called paycheck pronouns (Karttunen, 1976)). As annotation environment, the MMAX2 annotation software [5] was used.

To rule out data size as a possible explanation for performance shifts, datasets of equal size (about 30K) were randomly selected. The focus of the current experiments was on resolving identity and predicative relations. Table 1 gives some statistics about each dataset, such as the average sentence length and the number of coreferring NPs.

For all experiments we used an existing coreference resolver for Dutch, developed by Hoste (2005) and Hendrickx et al. (2008b). The system

[3] Website from CoNLL 2011: http://conll.bbn.com

[4] SoNaR is currently still under development.

[5] http://mmax2.net

follows a machine learning approach[6] based on the seminal work of Soon et al. (2001) and represents a mention-pair model. First, a classifier is trained to decide whether a pair of NPs is coreferential or not, after which coreference chains are built for the pairs of NPs that were classified as coreferential.

	#docs	#tokens	avg. senl	#coref NP
ADM	21	30,215	18.1	2,403
AUTO	15	30,058	14.6	2,411
EXT	29	29,940	15.9	2,381
INST	18	29,994	17.5	3,024
MED	213	30,001	14.4	1,995
JOUR	52	30,002	18.2	2,472
WIKI	15	30,340	18.9	3,480
DUO	56	29,740	19.7	3,063

Table 1: Size and number of coreferring NPs per dataset

All datasets were preprocessed in the same way. Tokenisation, lemmatisation, Part-of-Speech tagging and grammatical relations were based on the manually verified output of the Alpino parser (Bouma et al., 2001), i.e. gold standard dependency structures. For the DuOMAn data, however, no gold standard dependency trees were available. Named entity recognition was performed using MBT (Daelemans et al., 2003), trained on the 2002 CoNNL shared task Dutch dataset (Tjong Kim Sang, 2002) and an additional gazetteer lookup. As features we employ string matching, distance between sentences and NPs, grammatical role and named entity overlap, synonym/hypernym lookup using Cornetto (a Dutch database combining Dutch Wordnet (Vossen, 1998) and the Referentie Bestand Nederlands (Martin and Ploeger, 1999)) and local context. All instances were built between NP pairs going 20 sentences back in context. NPs that are not part of a coreferential chain (*singletons*) are included as negative examples. For more information we refer to Hoste (2005) and Hendrickx et al. (2008a).

Since the focus of this study is on genre, we decided not to train on different NP types (pronouns, common nouns and proper names) individually.[7] For all experiments we used Timbl version 6.3 (Daelemans et al., 2010) with default parameter settings.

Our experimental results are evaluated using the four scoring metrics as implemented in the scoring script from the coreference resolution task from the SemEval-2010 competition (Recasens et al., 2010):

- The MUC scoring software (Vilain et al., 1995) counts the number of links between the coreferential elements in the text, and looks how many links are shared or not between the gold standard coreferential chains and the system predictions. As MUC concentrates on links, elements that are not part of a coreferential chain, entities that are only mentioned once (*singletons*), are not taken into account in this scoring method.

- The B-cubed measure (Bagga and Baldwin, 1998) does not consider mere links between elements, but takes into account the coreferential clusters of elements referring to the same entity. B-cubed computes for every individual element in the text the precision and recall by counting how many elements are in the true coreferential cluster and how many in the predicted coreferential cluster.

- The CEAF measure (Luo and Zitouni, 2005) focuses on a one-to-one mapping of elements in the true and predicted coreferential clusters. Both B-cubed and CEAF measures are sensitive to the presence of many singletons, the larger the percentage of singletons, the higher these scores become (Recasens and Hovy, 2011).

- Recently, the BLANC measure (Recasens and Hovy, 2011) was developed to overcome problems with the other scoring methods. This measure is a variant of the Rand Index (Rand, 1971) adapted for coreference resolution and it averages over a score for correctly detecting singletons, and a score for detecting the correct cluster for coreferential elements.

An important remark to make here is that our system does not take into account chains of only one element. As a consequence, contrary to the SemEval-2010 competition, when we compute

[6] For an extensive overview of the different machine learning approaches for coreference resolution, we refer to the surveys of Ng (2010) and Poesio et al. (forthcoming)

[7] Hoste (2005) built a separate learning module for each of these NP types based on the motivation that the impact of different information sources varies per NP type.

	TRAIN	TEST
1.	one genre	that genre
	all genres but one	left out genre
	all genres	one genre
2.	one genre	other genres
3.	all LOO outliers	one genre

Table 2: Three sets of experiments

our scoring metrics, a singleton that is erroneously classified as part of a coreference chain is counted as an error. When it is correctly classified as a singleton, however, this is not represented in the scores.

In order to test cross-genre portability, we ran three sets of experiments (Table 2):

1. In the first set of experiments, we wanted to investigate whether adding more data is beneficial for the classifier. We trained the classifier on each genre individually and compared performance with different training set sizes. Three experiments were conducted: we first trained on each individual genre and tested on the relevant genre using ten-fold cross validation (each fold 27K vs. 3K). In a second experiment, the classifier was trained on all genres except one and tested on the one that was left out (210K vs. 30K). In a third experiment, we used all data, including genre-specific training material for training the classifier, in a ten-fold cross validation set-up (each fold 237K vs. 3K).

2. In a second set of experiments, we focused on the actual cross-domain portability. In order to test this, we each time trained on one genre and tested the performance of the classifier for each of the other genres.

3. Based on the results obtained in the second batch of experiments, we investigated whether some particular genres actually decrease performance when training on all data. In other words, does excluding outlier genres from training data increase performance? This was done by each time leaving out the worst-performing cross-domain genres and performing ten-fold cross validation.

3 Results

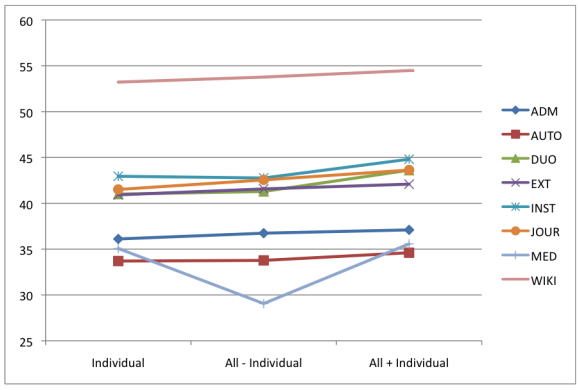
The results of the first round of experiments are presented in Figure 1. The dots marked as *individual* present the experiments in which each classifier was trained and tested on the same material. The scores for *All-individual* present experiments in which the classifiers are trained on a large and diverse training set of all different genres except the genre that is held out as a test set. The last experiments in the graph *All+individual* show the result when training on all genres including the held-out genre. Though the B-cubed and CEAF scores are lower than MUC, they present the same tendency: adding more and diverse training material improves performance, especially when genre-specific information is also included.[8] BLANC, however, seems to contradict the other metrics. Though the scores are higher, they reveal that larger training data proves only beneficial for three genres: INST, JOUR and MED. BLANC thus suggests that training only on in-domain material of some genre is the best approach.

This brings us to the cross-genre experiments, where we each time train on one genre and test on all the other genres individually until all genres have been once used as training data.[9] In order to represent the results, we ranked the classifier performance on each genre, ranging from the genre-classifier which on average performs worst when being applied to the other genres to the one performing best. We performed this ranking for each of the four evaluation metrics. The final ranking is visualized in Table 3. Although there are some differences between the metrics -we again observe that BLANC tends to differ more from the others - they all seem to agree that MED (medical text), DUO (unedited text) and INST (instructive text) constitute poor cross-genre training material. JOUR has been selected by MUC, B3 and CEAF as the best material for training on other genres. As we mentioned in Section 1 that most of the currently available datasets annotated with coreferential information consist of newspaper text, this result shows that this might indeed be a good choice.

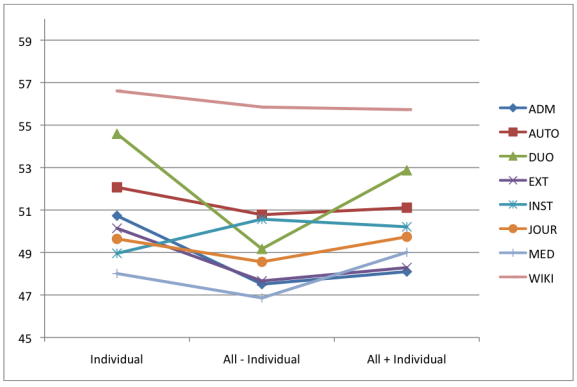
The four metrics confirmed that three genres had less generalization power, viz. MED, DUO and INST. In the third experiment, we aim to op-

[8] Because of space constraints we only incorporated two graphs in this paper.

[9] Train on ADM = test on AUTO; train on ADM test on DUO;....



(a) MUC F-measure



(b) BLANC F-measure

Figure 1: Performance comparison for each genre when training only on the genre, all the other genres, or both, respectively

MUC	B3	CEAF	BLANC
MED	MED	MED	MED
DUO	DUO	DUO	DUO
INST	INST	INST	INST
EXT	EXT	EXT	JOUR
WIKI	AUTO	AUTO	ADM
AUTO	ADM	ADM	AUTO
ADM	WIKI	WIKI	EXT
JOUR	JOUR	JOUR	WIKI

Table 3: Comparison of the worst (top) to best-performing (bottom) cross-domain genres per metric.

timize our selection of training data to get the best possible general performance. We hypothesize that leaving out those genres with less predictive power for other genres from the training material will increase overall performance. In this set of experiments we train on all data, including genre-specific information, and test on one genre while progressively leaving out those three genres. The results of this *reversed learning curve* for all metrics can be found in Table 4. Whenever a score is printed in bold, it is the best score obtained for a particular genre.

It is difficult to compare the different metrics with each other. We observe that only the BLANC metric confirms our expectation that the results are almost always better when poor training material is excluded from training. The results as measured with the other 3 metrics, however, show that leaving out data is only beneficial for half of the datasets. Overall, these results do not strongly confirm our hypothesis. An important observation to make is that, for all metrics, the performance gains which are obtained by leaving out data are modest, the effect of removing data is very small. Based on these observations we conclude that to get good generalization performance it is more important to have a large training set than to put time and effort in the composition of this training set.

3.1 Error Analysis

Three genres, viz. MED, DUO and INST, did not score high in the cross-domain experiments and were the first genres to be left out in the final experiments. An error analysis on this data imposed itself. Looking at the data itself we see that MED includes data of a scientific nature consisting of various entries in a medical encyclopedia. DUO contains mostly user-generated text as it consists of texts from blogs and newspaper articles together with a large set of reader comments. This type of data is rather different from the other genres as it is unedited, subjective, informal and more similar to spoken language than the other genres. INST contains various patient information leaflets and manuals in which exactly the same sentences are often repeated with only one word – mostly the name of the product – different. The above observations already hint at the low generalizability of these three genres.

Compared to the other genres, who on average contain 25% of coreferential NPs, we note that MED and INST contain a high number of coreferential NPs (respectively 33% and 37%) and DUO a rather low amount (viz. 18%). Looking at the data statistics given in Table 1, we observe that MED slightly differs from the others: it consists

Train \ Test	ADM	AUTO	DUO	EXT	INST	JOUR	MED	WIKI
MUC								
ALL	37.10	34.61	**43.61**	42.09	**44.81**	43.63	**35.57**	**54.48**
1MinMED	37.26	34.41	43.56	42.01	44.61	44.03		54.07
2MinDUO	**37.39**	**34.85**		**42.29**	44.51	**44.56**	35.44	54.35
3MinINST	37.06	34.00	31.02	41.81		44.46	34.72	54.21
B-cubed								
ALL	27.83	**29.77**	31.45	**30.64**	**31.66**	31.23	**26.08**	**30.84**
1MinMED	27.74	29.64	**31.68**	30.18	**31.66**	31.34		30.46
2MinDUO	**28.02**	29.46		30.11	31.26	**31.81**	25.99	30.58
3MinINST	27.87	29.54	31.02	30.01		31.61	25.18	30.64
CEAF								
ALL	29.48	**30.61**	29.79	**31.36**	28.42	31.42	**29.49**	26.31
1MinMED	29.11	30.33	29.96	30.26	**28.47**	30.86		**26.40**
2MinDUO	**29.73**	29.51		30.09	28.12	**31.62**	29.33	25.99
3MinINST	29.58	30.48	**22.97**	29.16		30.93	28.20	25.14
BLANC								
ALL	48.10	51.11	52.87	48.29	50.21	49.74	**49.01**	55.73
1MinMED	48.49	51.37	**54.70**	48.51	50.72	49.55		**56.66**
2MinDUO	48.73	51.49		48.73	**51.01**	**50.37**	48.15	56.11
3MinINST	**49.71**	**51.59**	54.16	**50.88**		49.61	48.49	56.17

Table 4: Results of the third set of experiments for all metrics and in comparison with training on all data.

of 213 smaller documents and the average sentence length is shorter, viz. 14.4 words. Moreover, looking at the subdivision of NPs we see that MED contains a large number of common nouns (89%) and only few pronouns (5%) and proper nouns (6%). In the other five datasets, this division ranges between 70-75% common nouns and 10-15% pronouns and proper nouns. When using MED as training data this results in a higher number of introduced errors between common nouns. Especially when no string matching features are found between two common nouns the resolver has a lot of difficulty into correctly classifying them. Of all genres we see that with MED pronouns and proper nouns are harder to recognize, which can be explained by their low coverage in the training data. Having a closer look at the DUO dataset, we see that the division between common, proper and pronouns is 64% - 14% - 22% – which is a high number of pronouns. Counterintuitively, this does not mean that resolving pronouns goes better when training on DUO. On the contrary,we see that although the resolution of pronouns rises slightly, more errors are introduced. Dutch pronouns also turned out to be difficult to resolve according to Hoste (2005) because of the inability to distinguish between anaphoric and pleonastic pronouns. The NP subdivision in INST is comparable to the five other genres, with a small preference for proper nouns. The high amount of reoccurring sentences in the data is also reflected in the features, the INST dataset scored best when performing in-domain experiments because of the many exact matches. Furthermore, as many technical NPs are not covered by WordNet (and these semantic features are crucial for most genres), important links between two NPs are missed.

In sum, these three genres have very specific features that seem to make them less predictive for other genres.

4 Conclusion

In this paper we explored the portability of an existing coreference resolver for Dutch when applied to eight different text genres: administrative texts, autocues, texts used for external communication, instructive texts, journalistic texts, medical texts, wikipedia and unedited new media texts. By comparing the performance on three sets of experiments, we found that larger training

set size improves performance, especially when genre-specific training material (10%) is included. We saw that excluding poor cross-genre training material does not always results in better scores neither can a drop in performance be perceived. This might imply that training on more data with higher predictive power is more important than training on various text genres. This is something we definitely wish to look into in closer detail in future work. Moreover, we would like to find out how much genre-specific training data is exactly needed to optimize performance. We discovered that especially genres containing very specific (e.g. scientific or unedited) data and having a different subdivision between pronouns, common and proper nous are less equipped for cross-genre experiments and thus have less generalization power.

We also observe that the different evaluation metrics for coreference research in use today, (MUC, B-cubed, CEAF and BLANC) tend to contradict each other and as a consequence hamper interpretation. This is a well-known problem within the community for which no solution has been found yet. In order to allow for a better comparison with the SemEval-2010 competition we intend to have a closer look at the effect of also scoring *singletons*.

Acknowledgments

The work presented in this paper was made possible by the STEVIN programme of the Dutch Language Union within the framework of the SoNaR project under grant number STE07014 and the Portuguese Science Foundation, FCT (Fundação para a Ciência e a Tecnologia). We would like to thank the anonymous reviewers for their helpful comments and valuable suggestions.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference*, pages 563–566.

Gosse Bouma and Geert Kloostermans. 2007. Mining Syntactically Annotated Corpora using XQuery. In *Proceedings of the Linguistic Annotation Workshop (held in conjunction with ACL 2007)*, pages 17–24, Prague, Czech Republic.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide coverage computational analysis of dutch. In *Computational Linguistics in the Netherlands 2000: selected papers from the twentieth CLIN meeting*.

Gosse Bouma, Walter Daelemans, Iris Hendrickx, Véronique Hoste, and Anne-Marie Mineur. 2007. The COREA-project, manual for the annotation of coreference in Dutch texts. Technical report, University Groningen.

Walter Daelemans, Jakub Zavrel, Antal van den Bosch, and Ko van der Sloot. 2003. MBT: Memory Based Tagger, version 2.0, Reference Guide. Technical Report ILK Research Group Technical Report Series no. 03-13, Tilburg University.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal van den Bosch. 2010. TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide. Technical Report ILK Research Group Technical Report Series no. 10-01, Tilburg University.

Hal Daumé III, Tejaswini Deoskar, David McClosky, Barbara Plank, and Jörg Tiedemann, editors. 2010. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, Uppsala, Sweden, July.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In *Proceedings of LREC 2004*, pages 837–840, Lisbon, Portugal.

Caroline Gasperin and Ted Briscoe. 2008. Statistical anaphora resolution in biomedical texts. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 257–264, Manchester, UK, August. Coling 2008 Organizing Committee.

Souha Hammami, Lamia Belguith, and Abdelmajid Ben Hamadou. 2009. Arabic anaphora resolution: Corpora annotation with coreferential links. *The International Arab Journal of Information Technology*, 6(5):481–489.

Iris Hendrickx and Véronique Hoste. 2009. Coreference Resolution on Blogs and Commented News. In *Anaphora Processing and Applications, Lecture Notes in Artificial Intelligence*, volume 5847, pages 43–53, Heidelberg.

Iris Hendrickx, Gosse Bouma, Frederik Coppens, Walter Daelemans, Véronique Hoste, Geert Kloosterman, Anne-Marie. Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008a. A coreference corpus and resolution system for Dutch. In *Proceedings of LREC 2008*, pages 144–149, Marrakech, Morocco.

Iris Hendrickx, Véronique Hoste, and Walter Daelemans. 2008b. Semantic and Syntactic features for Anaphora Resolution for Dutch. In *Proceedings of*

the 9th International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, volume 4919, pages 351–361, Haifa, Israel.

Véronique Hoste. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

Lauri Karttunen. 1976. Discourse referents. *Syntax and Semantics*, 7.

Lucie Kučová and Eva Hajičová. 2004. Coreferential relations in the Prague Dependency Treebank. In *Proceedings of DAARC 2004*, pages 97–102, Azores, Portugal.

Laurence Longo and Amalia Todirascu. 2010. Genre-based reference chains identification for french. *Investigationes Linguisticae*, 21:57–75.

Xiaoqiang Luo and Imed Zitouni. 2005. Multi-lingual coreference resolution with syntactic features. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 660–667.

Bernardo Magnini, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi-Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of LREC 2006*, pages 963–968, Genoa, Italy.

Willy Martin and Jeannette Ploeger. 1999. Tweetalige woordenboeken voor het Nederlands: het beleid van de Commissie Lexicografische Vertaalvoorzieningen. *Neerlandica Extra Muros*, 37:22–32.

Vincent Ng. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411.

Jin-Dong Kim Ngan Nguyen and Junichi Tsujii. 2008. Challenges in pronoun resolution system for biomedical text. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC 2008*, pages 1170–1174, Marrakech, Morocco.

Massimo Poesio, Simone Paolo Ponzetto, and Yannick Versley. forthcoming. Computational models of anaphora resolution: A survey. *Linguistic Issues in Language Technology*.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering,*.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Marta Recasens, Lluíz Márquez, Emili Sapena, M. Antònia Martí, Mariona Tauleé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 Task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pages 1–8, Uppsala, Sweden.

Marta Recasens. 2010. *Coreference: Theory, Annotation, Resolution and Evaluation*. Ph.D. thesis, Department of Linguistics, University of Barcelona, Barcelona, Spain, September.

Kepa Joseba Rodríguez, Franceska Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of Wikipedia and blogs in the Live Memories Corpus. In *Proceedings of LREC 2010*, pages 157–163, Valletta, Malta.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. In *Proceedings of LREC 2010*, pages 2471–2477, Valletta, Malta.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 155–158, Taipei, Taiwan.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52.

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston, 2009. *OntoNotes Release 3.0. LDC2009T24*. Linguistic Data Consortium.

Xiaofeng Yang, Jian Su, GuoDong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *Proceedings of Coling 2004*, pages 226–232, Geneva, Switzerland, Aug 23–Aug 27.