

Towards a Balanced Named Entity Corpus for Dutch

Bart Desmet^{1,2} and Véronique Hoste^{1,2}

¹LT3, Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium

²Department of Applied Mathematics and Computer Science, Ghent University
Krijgslaan 281 (S9), 9000 Gent, Belgium
bart.desmet, veronique.hoste@hogent.be

Abstract

This paper introduces a new named entity corpus for Dutch. State-of-the-art named entity recognition systems require a substantial annotated corpus to be trained on. Such corpora exist for English, but not for Dutch. The STEVIN-funded SoNaR project aims to produce a diverse 500-million-word reference corpus of written Dutch, with four semantic annotation layers: named entities, coreference relations, semantic roles and spatiotemporal expressions. A 1-million-word subset will be manually corrected. Named entity annotation guidelines for Dutch were developed, adapted from the MUC and ACE guidelines. Adaptations include the annotation of products and events, the classification into subtypes, and the markup of metonymic usage. Inter-annotator agreement experiments were conducted to corroborate the reliability of the guidelines, which yielded satisfactory results (Kappa scores above 0.90). We are building a NER system, trained on the 1-million-word subcorpus, to automatically classify the remainder of the SoNaR corpus. To this end, experiments with various classification algorithms (MBL, SVM, CRF) and features have been carried out and evaluated.

1. Introduction

Named Entity Recognition (NER) is the task of automatically identifying names within text and classifying them into categories, such as persons, locations and organizations. NER started as an information extraction subtask, but has since evolved into a distinct task essential for information retrieval, question answering, and as a preprocessing step for coreference resolution and various other NLP problems.

An extensive literature on the subject exists (Chinchor, 1998; Tjong Kim Sang and De Meulder, 2003; Cucerzan, 2007), with NER approaches roughly falling into three categories: hand-crafted, machine learning and hybrid systems. Hand-crafted approaches require manual rule creation, a time-consuming process which hinders easy porting to new domains or languages. Supervised machine learning solutions, on the other hand, rely on an annotated training corpus to infer patterns associated with named entities, based on morphological, syntactic, lexical and contextual features. Hybrid systems combine both approaches. Such systems are in widespread use and have proven their effectiveness, with Zhou and Su (2002) reporting near-human performance on English data.

The bottleneck for the development of machine learning applications is its dependence on, preferably large, annotated training corpora. Among the named entity resources for English are the manually annotated data sets from the MUC-7 Named Entity Task (Chinchor (1998), 162,692 tokens) and the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder (2003), 301,418 tokens), and the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), which provides a named entity and coreference annotation layer for the Penn Treebank corpus of Wall Street Journal texts (1,173,766 tokens). For Dutch, however, the data from the CoNLL-2002 shared task (Tjong Kim Sang, 2002a), containing 309,686 tokens from four

editions of the Belgian newspaper "De Morgen" of 2000, constitute the only corpus annotated with named entity information that is readily available at present.

The pressing need for a substantial corpus of Dutch text, not only for NER, is addressed in the STEVIN¹-funded SoNaR project². It aims to produce a 500-million-word reference corpus of written Dutch containing a wide spectrum of genres and text types (Oostdijk et al., 2008). A 1-million-word subset will be provided with a number of manually corrected annotation layers, including four semantic ones: named entities, coreference relations, semantic roles and spatiotemporal expressions (Schuurman et al., 2009). The subset contains these various text types, reflecting the global corpus design. This diversity, which was particularly lacking in the Dutch CoNLL-2002 data set, should allow for a more robust classifier and better cross-corpus performance. The manually annotated subcorpus will be used to train classifiers for the automatic annotation of the remaining 499 million words. These annotations will not be checked manually, but will be made available with confidence scores.

For the named entity annotation of the corpus, new annotation guidelines were developed, based on the guidelines from MUC-7 (Chinchor and Robinson, 1997) and ACE (LDC, 2008). A number of adaptations were made, most notably the addition of separate classes for products and events, and the annotation of metonymy.

In the remainder of this paper, we will discuss and motivate the annotation guidelines in Section 2, we present an evaluation of the guidelines based on inter-annotator agreement scores in Section 3 and give an overview of the use of the guidelines within the context of the SoNaR project in Section 4. In Section 5, we describe our experiments with a number of classifiers, and evaluate the generalization

¹<http://taalunieversum.org/taal/technologie/stevin/>

²<http://lands.let.ru.nl/projects/SoNaR/>

performance of the best classifier on the annotated corpus. Section 6 concludes this paper.

2. Annotation guidelines

The SoNaR named entity annotation guidelines³ are based on the MUC-7 and ACE annotation schemes for English named entities. The aim of the new guidelines was to achieve consistent and fine-grained annotation of Dutch text. To that end, the guidelines describe the delimitation of named entities (see 2.1.), the classification into main types (2.2.) and subtypes (2.3.), and the markup of metonymic usage (2.4.). For an overview of the possible annotations, see Figure 1.

2.1. Span

Named entities can be defined as “unique identifiers of referents in reality”, such as proper names (Chinchor and Robinson, 1997; Borrega et al., 2007). In the context of NER, this definition is often interpreted broadly to include temporal (e.g. dates, times) and numerical (e.g. monetary values, percentages) expressions. This was deemed unnecessary for our annotation scheme, in part because of the presence of a dedicated spatiotemporal annotation layer. In practice, it is often unclear whether a given phrase should be considered a unique identifier or not. This is especially true for named entities of the types product, event and miscellaneous (see 2.2.). Consider the following example:

- (1) Koning Albert II zal de twee Koninklijke Besluiten van minister van Werk Peter Vanvelthoven (sp.a) niet ondertekenen.
English: King Albert II will not sign the two Royal Decrees by the minister of Employment, Peter Vanvelthoven (sp.a).

In sentence 1, it is debatable whether “King”, “Royal Decrees” and “Employment” should be considered (part of) a unique identifier. For reasons of consistency, a pragmatic approach was taken to the delimitation of named entities. All words starting with a capital letter that are not the first word of a sentence and are not entirely capitalized, are taken to be named entities. All sentence-initial, uncapitalized or all caps words that can unequivocally be considered unique identifiers are annotated as well. Sentence 1 will thus be annotated as follows:

- (2) Koning [Albert II] zal de twee [Koninklijke Besluiten] van minister van [Werk] [Peter Vanvelthoven] ([sp.a]) niet ondertekenen.
English: King [Albert II] will not sign the two [Royal Decrees] by the minister of [Employment], [Peter Vanvelthoven] ([sp.a]). (sp.a is a Belgian political party)

Named entities can be part of a word that as a whole is not a named entity, e.g. “London-based”. In English, such structures are rare and will often be annotated fully as MISC, or not at all (Nothman et al., 2009). Given the frequency of concatenated compounds in Dutch, we chose to annotate named entities word-internally:

- (3) [Apple]topman [Steve Jobs] kondigde het [iPhone]platform op [Macworld 2007] aan.
English: [Apple] [CEO] [Steve Jobs] announced the [iPhone] platform at [Macworld 2007].

2.2. Main types

Tokens marked as named entities can be classified as one of six main types:

- PER: names of persons, fictitious characters (human or otherwise), gods, artist names and generational suffixes; e.g. Elizabeth Bennet, Lassy, Vishnu, Sting, George Bush Sr.
- ORG: names of organizations, including organizational suffixes; e.g. European Parliament, Google Inc.
- LOC: names of locations, and derived adjectives; e.g. Paris, Mount Everest, Japanese
- PRO: names of products, awards, works of art and languages; e.g. Office 2007, Academy Award, Pride and Prejudice, Sanskrit
- EVE: names of events; e.g. World War II, Katrina
- MISC: miscellaneous names; e.g. Mesozoic, Employment

PER, ORG and LOC are the usual suspects in named entity annotation, with MISC sometimes acting as a backup class. We added the PRO and EVE categories to obtain good coverage of possible named entities and to allow for consistent metonymy roles (see 2.4.). The EVE category is only used for events that are designated by a named entity, and not for relations between named entities. The MISC category was reserved for instances produced by the broad definition of span (2.1.) that fitted more than one, or none of the five other main types.

2.3. Subtypes

For four of the six main classes, mutually exclusive subtypes are to be annotated. Named entities of type person and miscellaneous receive no subtype annotation. Organizations are classified as governmental, commercial or miscellaneous. For locations, nine subtypes are distinguished: continents, countries, regions (such as provinces and natural regions), population centres (such as cities and neighbourhoods), lines (such as streets and highways), points (such as buildings and parks), water bodies (such as seas and rivers), extraterrestrial locations (such as planets and galaxies) and fictional locations. Products are classified as shares (on the stock market), languages or miscellaneous, and events are labeled as either human or natural.

The motivation for including subtypes was twofold:

1. It allows for fine-grained annotation, as required for question answering tasks, without compromising a robust, coarse-grained main type structure.
2. It provides useful information for the classification of usage (2.4.) and for the other semantic annotation layers in the SoNaR project.

³<http://lt3.hogent.be/sonar/share/>

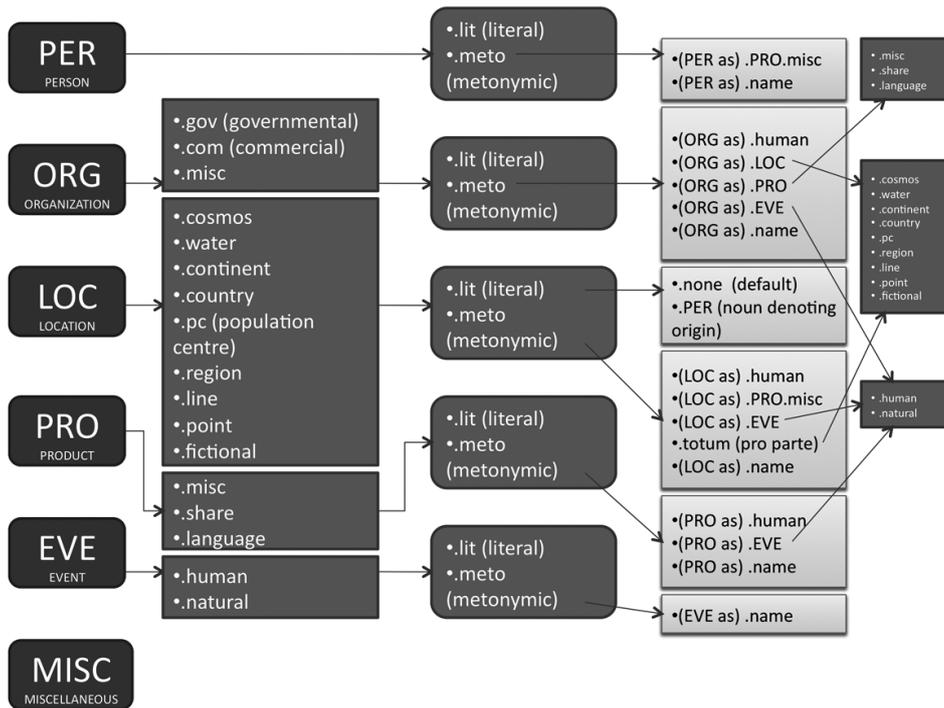


Figure 1: Annotation scheme for named entities, with categories for main type, subtype, usage and metonymic roles.

The markables in Sentence 3 would be classified as follows:

[Apple]_{ORG.com}topman [Steve Jobs]_{PER} kondigde het [iPhone]_{PRO.misc}platform op [Macworld 2007]_{EVE.human} aan.
 English: [Apple]_{ORG.com} [CEO]_{MISC} [Steve Jobs]_{PER} announced the [iPhone]_{PRO.misc} platform at [Macworld 2007]_{EVE.human}.

2.4. Usage

An important issue we wanted to address in our annotation scheme was the metonymic use of named entities. Consider Sentence 4:

- (4) Het [Witte Huis] koos voor moderne werken, waaronder een [Rothko].
 English: The [White House] opted for modern works of art, including a [Rothko].

Cases like “White House” being classified as LOC rather than ORG are a common mistake (Nothman et al., 2009). By marking whether a named entity is used literally (.lit) or metonymically (.meto), we can consistently label named entities for their *literal* main type, and use metonymic roles to point to their *intended* main type (PER, ORG, LOC, PRO or EVE). This approach was inspired by Markert and Nissim (2002) and (2007).

Because it is often impracticable to determine whether a named entity is used metonymically as PER or as ORG, we combined them in the intended type “human” (see for example 5, where “White House” might refer to a person, namely the U.S. president, or to an organization-like group of people such as the White House staff).

When a name is used as a mere signifier, the intended type is “name”.

- (5) Het [Witte Huis]_{LOC.point.meto.human} koos voor moderne werken, waaronder een [Rothko]_{PER.meto.PRO.misc}.
 English: The [White House]_{LOC.point.meto.human} opted for modern works of art, including a [Rothko]_{PER.meto.PRO.misc}.

Marking metonymy does not only do away with confusable main types, it should also benefit the automatic annotation of other semantic layers. For example, a coreferential resolution algorithm could link an inanimate noun phrase like “the painting” to “Rothko” in Sentence 5 if it has access to named entity classifier output that does not only mark “Rothko” literally as an (animate) person, but also metonymically as a product.

3. Guideline evaluation

In order to evaluate the guidelines, two linguists annotated a set of eight randomly selected texts from the corpus, containing 14,244 tokens in total. Two evaluation metrics were used: Kappa (Carletta, 1996) and F-score ($\beta = 1$) (Van Rijbergen, 1979). F-scores were calculated by taking one annotator as the gold standard and scoring the annotations of the other for precision and recall. This yields the same results as averaging the precision or the recall scores of both annotators, when using the other as a gold standard.

Scores were calculated on 5 levels: span, main type, subtype, usage and metonymic role. For each level, scores were calculated on the entire set, and on a subset containing only those tokens (i) on which both annotators agreed on the *preceding* level, and (ii) which can receive annotation on the *current* level (MISC and PER, for example, are not included in the subset for subtype, because they cannot

receive subtype annotation). The results can be found in Table 1, in which absolute counts for span, main type and usage are also included.

The results show high agreement scores for all levels, most notably span. However, it is difficult to evaluate the agreement for minority classes by using these global metrics. For the subtype level, for example, rare classes with low inter-annotator agreement would hardly influence the overall subtype agreement score.

4. Annotation implementation

The SoNaR corpus will comprise a wide variety of texts, including traditional text types (such as newswire, manuals, autocues, fiction and reports) as well as new media (such as blogs, forums, chat and SMS), for a total of 500 million words. A representatively diverse 1-million-word subset is being annotated manually, and will serve as the gold standard for the automatic annotation of the entire corpus. This diversity is essential to training automatic classifiers that can be applied to the corpus as a whole, and should also make it an interesting corpus for research on domain adaptation (Nothman et al., 2009).

Manual annotation is done using the MMAX2 annotation tool (Muller and Strube, 2006). For the named entity task, six annotation layers were created - one per main type. Per text, each annotation layer is stored as a standoff XML file, the content of which is defined by a scheme file. These scheme files contain the possible attributes for every annotation, such as the available subtypes, the choice between literal and metonymic usage, and the metonymic role, if applicable.

The scheme files also provide the possibility to mark an annotation as uncertain with respect to span, main type, subtype or metonymic usage. This information can be used to clean the corpus of unwanted noise, which can negatively influence classifier performance.

Annotation speed averaged around 3,500 words per hour. Taking into account the verification of the annotations by a second annotator, the actual annotation speed will be closer to 2,000 words per hour.

5. Automatic classification

5.1. Experiments

In order to develop an automatic NER system, we experimented with a number of machine learning algorithms. The initial experiments were carried out on the Dutch data set of the CoNLL-2002 shared task, which is the only Dutch resource on which named entity classification results are available for comparison (Tjong Kim Sang, 2002a; Bogers, 2004). The data consist of a training set (`ned.train`, 218,737 lines), a development set for finetuning the system (`ned.testa`, 40,656 lines) and a test set (`ned.testb`, 74,189 lines). Results on the test set were evaluated with the provided `conlleval` Perl script.

Three machine learning implementations were used for the experiments:

- TiMBL, version 6.2.1 (Daelemans et al., 2009), a memory-based learner (MBL), used with the IB1 algorithm, inverse linear distance, the Jeffrey divergence metric and $k = 7$;

- CRF++⁴, version 0.53, a sequence tagger using Conditional Random Fields (CRF), used with standard parameters;
- YamCha, version 0.33 (Kudo and Matsumoto, 2003), a sequence tagger using Support Vector Machines (SVM), used with standard parameters.

The following information was encoded in a feature vector for every token:

- The token itself;
- The length of the token;
- Prefixes and suffixes of the token, up to length 4;
- POS and chunk tags, obtained by preprocessing the data with a memory-based shallow parser (Daelemans and van den Bosch, 2005);
- Morphological information, encoded as binary features indicating whether the token starts with or contains a capital letter, is completely in upper or lower case, and whether it contains or is entirely composed of digits, punctuation characters or hyphens. This information is also merged in a symbolic word shape feature;
- External information, namely a feature indicating whether the token is present in an external list of function words;
- Pattern features, indicating whether the token matches regular expressions for initials and URLs;
- Contextual information, such as a binary feature indicating whether the token is in sentence initial position. Contextual information was also provided by windowing the focus token with 3 tokens and their feature vectors to the left and 1 to the right.

5.2. Results on the CoNLL data

The three systems were trained and refined using the CoNLL train and development set. Table 2 presents the results per classifier, along with the results of two participants in the CoNLL-2002 shared task competition, namely the best overall classifier (Carreras et al., 2002) and the best-performing memory-based system (Tjong Kim Sang, 2002b), and the results obtained in (Bogers, 2004), also using memory-based learning.

The MBL approach beats the results of the best-performing memory-based classifier by Tjong Kim Sang, which was, however, implemented as a language-independent system. The classifier by Bogers, which was built specifically for Dutch, is outperformed by a small margin.

The best results are achieved with the SVM-based system. It improves the precision, recall and F-scores of the system by Carreras et al., which made use of binary AdaBoost classifiers, by 1.74, 1.89 and 1.82 percentage points, respectively.

⁴<http://crfpp.sourceforge.net/>

Level	Total set		Subset			
	Kappa	$F_{\beta=1}$	Kappa	$F_{\beta=1}$	Tokens	Distribution
Span	0.97	99.62	0.97	99.62	14244	13293 non-NE, 897 NE, 54 NA
Main type	0.94	99.23	0.92	93.76	897	150 PER, 225 ORG, 241 LOC, 115 PRO, 62 EVE, 48 MISC, 56 NA
Subtype	0.92	99.12	0.94	97.67	643	32 NA
Usage	0.91	98.93	0.93	94.58	793	733 literal, 17 metonymic, 43 NA
Role	0.91	98.90	1.00	100.00	17	0 NA

Table 1: Inter-annotator agreement scores per level, token count and distribution (NA = no agreement).

	Precision	Recall	$F_{\beta=1}$
MBL	71.10	73.28	72.17
CRF	77.71	73.94	75.78
SVM	79.57	78.18	78.87
Tjong Kim Sang	72.56	68.88	70.67
Carreras et al.	77.83	76.29	77.05
Bogers			71.90

Table 2: Precision, recall and F-scores on the CoNLL-2002 data set

The significance of these classification results is limited, considering that the CoNLL shared task was a language-independent competition, and the state of the art has advanced since 2002. However, the named entity classifiers tested on the CoNLL-2002 data set are the only ones with which comparison is possible for Dutch. We therefore conclude that our best-performing system is adequate for testing the generalization performance on the new SoNaR corpus, and provides a reference point for future systems tested on it.

5.3. Results on the SoNaR data

We applied the SVM-based classifier with ten-fold cross validation on the section of the SoNaR 1 million word corpus that was available with annotations at the time of testing (702,846 tokens). The overall results and results per main type are presented in Table 3.

	Precision	Recall	$F_{\beta=1}$
Overall	76.41	74.33	75.35
PER	77.07	79.84	78.43
ORG	69.40	67.68	68.53
LOC	82.58	88.88	85.61
EVE	82.53	46.65	59.61
PRO	51.24	23.45	32.18
MISC	52.21	37.48	43.64

Table 3: Precision, recall and F-scores per main type using YamCha with ten-fold cross validation on the SoNaR corpus

The results show high performance overall and for the PER and LOC categories, and moderate performance for ORG. The F-scores for the EVE, PRO and MISC categories are lower, especially because of low recall scores. Table 4 gives a quantitative overview of the errors made across the main types.

In order to find patterns in the errors, a qualitative analysis was carried out on the ones that caused the low recall and precision scores for PRO and MISC, and the low recall score for EVE. We found three causes that explain the majority of the errors, and formulate suggestions to improve the performance of the classifier.

5.3.1. Annotation

Due to inconsistent annotation, a number of errors in the PRO and MISC categories are in fact correctly labeled, but do not match with the incorrect gold standard annotation. The gold standard will be entirely checked to remove such errors. Improved consistency should also allow for easier generalization by the classifier.

5.3.2. Abbreviations

Abbreviations are often incorrectly labeled, because the system cannot link them to their corresponding full form, as in Sentence 6.

- (6) De politie zal de vermiste persoon signaleren in het [Nationaal Schengen Informatie Systeem]_{MISC} ([NSIS]_{ORG}).
English: The police will report the missing person in the [National Schengen Information System]_{MISC} ([NSIS]_{ORG}).

NSIS should be labeled as MISC instead of ORG, but the system fails to give the abbreviation the same label as its full form (which may appear anywhere in the left context of the document). An abbreviation resolution module should be implemented to inform the classifier of full forms.

5.3.3. External information

A large portion of the errors cannot easily be solved with text-internal information only. Consider the incorrectly labeled Sentence 7:

- (7) Welke voordelen bleek [Yentreve]_{PER} tijdens de studies te hebben?
English: Which advantages did [Yentreve]_{PER} appear to have during the trials?

Yentreve is a pharmaceutical product, not a person. The system would benefit from access to external information sources to classify this instance correctly. One way of providing such information is by including gazetteer features. Gazetteers are lists of names pertaining to a certain category. For every gazetteer, a binary feature is used to indicate if the token appears in it. Another way of including external information would be to query Wikipedia. Kazama

gold/predict	PER	ORG	LOC	EVE	PRO	MISC	O	Total
PER	14929	293	953	1	165	165	1055	17561
ORG	1021	8101	1082	26	310	443	877	11860
LOC	770	612	18749	19	171	180	808	21309
EVE	74	237	265	833	51	70	99	1629
PRO	877	703	749	32	1348	312	2011	6032
MISC	345	920	605	54	270	1667	745	4606
O	396	394	373	21	467	294	637904	639849
Total	18412	11260	22776	986	2782	3131	643499	702846

Table 4: Contingency tabel of the errors made by the SVM system on the SoNaR data. Predicted labels are on top of the table, correct labels are on the left. Correctly predicted instances are in bold face.

and Torisawa (2007) describe a successful Wikipedia retrieval system for named entity recognition.

6. Conclusions and future work

In this paper, we presented the named entity annotation guidelines for the SoNaR project. Their aim is to provide consistent and fine-grained annotations that capture useful information for subsequent classification tasks and information retrieval. To this end, a pragmatic approach was taken for the delimitation of named entities, resulting in high inter-annotator agreement scores for span.

The annotation scheme also provides reliable guidelines for the annotation of subtypes and metonymic usage, making SoNaR the first publicly available corpus in the Dutch domain to incorporate such information.

The experiments carried out for the automatic recognition and classification of main types yielded satisfactory results, compared to the CoNLL-2002 shared task submissions. Additional research will be conducted to improve the performance on main types, by implementing abbreviation resolution, adding external information from gazetteers and Wikipedia and experimenting with classifier ensembles. Another area of future research is the automatic classification of subtypes and metonymy.

7. References

- T. Bogers. 2004. Dutch named entity recognition: Optimizing features, algorithms, and output. Master's thesis, Universiteit van Tilburg.
- O. Borrega, M. Taulé, and M.A. Marti. 2007. What do we mean when we speak about named entities? In *Proceedings of Corpus Linguistics*.
- J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.
- X. Carreras, L. Marquez, and L. Padro. 2002. Named entity extraction using Adaboost. In *Proceedings of CoNLL-2002*, Taipei, Taiwan.
- N. Chinchor and P. Robinson. 1997. MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.
- N. Chinchor. 1998. Overview of MUC-7. In *Proceedings of the 7th Message Understanding Conference*.
- S. Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL*, pages 708–716.
- W. Daelemans and A. van den Bosch. 2005. *Memory-based Language Processing*. Cambridge University Press.
- W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2009. TiMBL: Tilburg Memory Based Learner, version 6.2, reference guide. Technical Report 09-01, ILK Research Group.
- J. Kazama and K. Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707, Prague, Czech Republic, June.
- T. Kudo and Y. Matsumoto. 2003. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 24–31.
- LDC, 2008. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6*. Linguistic Data Consortium, Philadelphia, USA. <http://projects.ldc.upenn.edu/ace/>.
- K. Markert and M. Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. In *Proceedings of the Third International Language Resources and Evaluation (LREC'02)*, pages 1385–1392, Las Palmas, Spain.
- K. Markert and M. Nissim. 2007. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague, Czech Republic.
- C. Muller and M. Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt, Germany.
- J. Nothman, T. Murphy, and J.R. Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 612–620, Athens, Greece.
- N. Oostdijk, M. Reynaert, P. Monachesi, G. Van Noord, R. Ordeman, I. Schuurman, and V. Vandeghinste. 2008. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

- I. Schuurman, V. Hoste, and P. Monachesi. 2009. Cultivating trees: Adding several semantic layers to the Lassy treebank in SoNaR. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, The Netherlands.
- E.F. Tjong Kim Sang and F. De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 142–147, Edmonton, Canada.
- E.F. Tjong Kim Sang. 2002a. 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.
- E.F. Tjong Kim Sang. 2002b. Memory-based named entity recognition. In *Proceedings of CoNLL-2002*, Taipei, Taiwan.
- C.J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworth, London.
- R. Weischedel and A. Brunstein, 2005. *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia, USA.
- G.D. Zhou and J. Su. 2002. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480, Philadelphia, USA.