Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR

Ineke Schuurman
K.U.Leuven
Centrum voor Computerlinguïstiek
ineke.schuurman@ccl.kuleuven.be

Veronique Hoste
University College Ghent
LT3
veronique.hoste@hogent.be

Paola Monachesi
Utrecht University
Uil-OTS
paola.monachesi@phil.uu.nl

1 Introduction

Within the STEVIN[1] project Large Scale Syntactic Annotation of written Dutch (LASSY), a manually corrected treebank of 1 million words is constructed. Lassy is part of a series of annotation projects for modern written and spoken Dutch. More specifically, it is an extension of the D-Coi and CGN projects,[2] and constitutes the core of SoNaR, a 500 million words reference corpus of modern written Dutch.[3] One of the goals of the latter project is to enrich the corrected treebank produced in Lassy[4] with several semantic layers.

For a general overview of the relations between D-Coi, Lassy and SoNaR, cf [19]. In this paper we will concentrate on the semantic layers of SoNaR core: (1) named entity labeling, (2) annotation of co-reference relations, (3) semantic role labeling and (4) annotation of spatial and temporal relations. Of these (2) originates from the STEVIN-project COREA,[5] (3) and (4) from D-Coi, whereas (1) is a new area within STEVIN.

---

[1] Funded by both the Dutch and Flemish governments, the present joint Dutch-Flemish STEVIN programme was started in 2004 (http://taalunieversum.org/taal/technologie/stevin/)

[2] http://lands.let.ru.nl/projects/d-coi/; http://lands.let.kun.nl/cgn/ehome.htm

[3] http://lands.let.ru.nl/projects/SoNaR/. The first phase of SoNaR started January 2008.

[4] The remaining 499 million words of SoNaR will also be tagged and parsed, but there will be no manual correction.

[5] http://www.cnts.ua.ac.be/~hoste/corea.html

2   Four semantic layers

So far, the creation of semantically annotated corpora has lagged behind dramatically. Within the STEVIN-programme, four birds are now being killed with one stone, with four layers of semantic annotions being added to an existing, manually corrected treebank.
The layers will be added in the order in which they are presented below, this way the layers at the end can profit from the results obtained earlier.

2.1   Named Entity and Co-Reference Labeling

Factoid question answering and information extraction relies heavily on the detection of named entities[6] in texts, since names in general tend to be salient context words. If information is extracted from a text, it often involves answering questions of the type *who, what, where?*, all of which can have names as answers. However, much information about these entities is often hidden in anaphoric references. Because of the high frequency of anaphoric expressions, resolving them is important for text understanding.

Named entity labeling.   Lacking substantial Dutch corpora annotated with named entity information we develop annotation guidelines on the basis of a comparative study of existing guidelines such as the MUC[7] [4], ACE[8] and TIDES guidelines[9]. While MUC considered three entity types (person, organization, location), ACE further divides locations into geo-political entities and facilities and also adds weapons, substances, and vehicles. Our main focus is on the TIDES named entity annotation, which consists of entity names, temporal expressions, and number expressions. The expressions to be annotated are "unique identifiers" of entities (persons, locations, organizations), times (dates, times, and durations), and quantities (money, measures, percentages, and cardinal numbers). Several of these are further disambiguated when annotating spatiotemporal expressions, cf. section 2.3. We furthermore investigate the annotation of the metonymic use of named entities, e.g. BMW as organization which can be used to denote a car or an index on the stock market.
On the basis of the semantically enriched, one million word core treebank of SoNaR, at a later stage a combined classifier will be developed in which diverse classifiers (such as a maximum entropy learner, a rule learner and a memory-based

---

[6]Like names of persons, organizations, and locations or expressions of times and quantities
[7]http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html
[8]http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v5.6.1.pdf
[9]http://www.nist.gov/speech/tests/ie-er/er_99/er_99.htm

learner) are combined with different types of output classes (singular class labels versus trigrams) and under different conditions (as in [9]). We will investigate different types of features, including orthographic, syntactic and lexical features and we will include seed-list information. Unannotated data will be used to expand the seed lists (as in [3] and to enhance the instance or rule base [6]. The resulting classifier will be applied to the 499 million word corpus and will provide confidence scores in order to facilitate post-processing.

Co-reference annotation. In order to detect all information available on a given name or entity, co-reference resolution is needed to unveil all the information referring to this given entity. In the last decades, corpus-based techniques have become increasingly popular for the resolution of co-referential relations (e.g. [17], [24] for English). The use of a corpus-based methodology was enabled by the creation of co-referentially annotated corpora, such as those developed within the framework of the MUC-6 and MUC-7 Message Understanding Conferences [11]. A more recent annotation effort is the creation of the English ACE (Automatic Content Extraction)[10] data sets.

In order to allow for the development of a corpus-based system for Dutch, which is able to detect co-referential relations between different types of noun phrases, including named entities, definite and indefinite NPs and pronouns, a substantial annotated corpus is required. This has led to the creation of the KNACK-2002 corpus [13] and the creation of the first machine learning system for Dutch co-reference resolution [12]. The guidelines as used in the KNACK-2002 corpus, which were largely based on the MUC-6 and MUC-7 annotation scheme for English, were the basis for the broader annotation guidelines of the Stevin COREA ("Co-reference Resolution for Extracting Answers") project [10], which is applied in SoNaR. SoNaR will lead to an additional 1 million annotated tokens for Dutch, on top of the 325,000 tokens produced in the projects above.

The following co-referential relations are labeled:

1. **Identity or strict co-reference** as in "**Xavier Malisse** *heeft zich geplaatst voor de halve finale.* **De Vlaamse tennisser** *zal dan tennissen (...)"* (English: "Xavier Malisse qualified for the semifinals. The Flemish tennis player will play (...)")

2. **Time-indexed co-reference** as in "**Bert Degraeve**, *die tot voor kort* **gedelegeerd bestuurder van de VRT** *was, gaat aan de slag bij staaldraadproducent Bekaert als* **chief financial and administration manager."** *(English: "Bert Degraeve, until recently managing director of the VRT, starts to work*

---

[10]www.nist.gov/speech/tests/ace/

at steel wire producer Bekaert as chief financial and administration manager." In the example, the two job descriptions refer to the same person, Bert Degraeve, but at a different point in time.

3. **Type-token co-reference** as in "*Ik verkies* **de rode auto**, *maar mijn man wou* **de grijze**". *(English: "I prefer the red car, but my husband wanted the gray one.").* The example sentence talks about two distinct cars, a red one and a gray one. "de grijze" denotes something like an object type rather than an object token.

4. **Part/whole co-reference** as in "*Hij kon zijn auto niet starten. De benzinetank was leeg." (English: "He could not start the car. The gas tank was empty.")*

5. **Modality and negation** as in "**Filip Dewinter** *had wel eens* **de nieuwe burgemeester van Antwerpen** *kunnen worden." (English: "Filip Dewinter could have become the new major of Antwerp.")* The use of a modal verb cluster like "had wel eens kunnen worden" indicates a fuzzy relation between anaphor and antecedent.

6. **Predicate nominals** as in "**Vivendi Universal** *is* **de tweede sterkste stijger binnen de DJ Stoxx50".** *(English: "Vivendi Universal is the second largest performer in the DJ Stoxx50.")* Although justly criticized by [5] and [26], we believe that the annotation of co-referential relations between predicate nominals is useful from an application point of view (e.g. information extraction).

7. **Appositions** as in "**Jones**, **de voorzitter van de raad van bestuur**, *verminderde zijn belang met 0,2%." (English: "Jones, the president of the Board of Directors, reduced its interests by 0.2%").* A special case of co-reference is that of co-referential appositions. Appositions come in two flavours: repetitive and restrictive.

8. **Bound anaphora** as in "**Niemand** *verliest graag* **zijn** *job." (English: "Nobody likes to lose his job".)* Rather than making statements about singular objects in the real world, the preceding example expresses properties of general categories.

9. **Metonymy** as in "**Laken** *heeft gedurende de hele periode 1960-1961* **zijn** *rol in de coulissen gespeeld." (English: "During the whole period 1960-1961 Laken has played its role behind the scenes.)* in which the castle of king Baudoin, Laken, is a common description of the Belgian monarchy.

In the COREA project, the co-reference annotations were based on the annotated corpora of the D-Coi project, which provide ample syntactic information. This same approach is used in SoNaR. The annotation process is accelerated through the application of the automatic co-reference resolution system developed in the COREA project, based on [12], which is retrained during the annotation process.

2.2 Semantic Role Labeling

During the last few years, corpora enriched with semantic role information have received much attention, since they offer rich data both for empirical investigations in lexical semantics and large-scale lexical acquisition for NLP and Semantic Web applications. Several initiatives have emerged at the international level to develop annotation systems of argument structure. Two projects have a leading position in this area, namely FrameNet [14] and PropBank [15] with their own semantic annotation schemes. Of these two, it was decided in D-Coi to annotate the corpus with semantic roles according to the PropBank approach as it provides fewer, more syntactically driven argument labels than FraMeNet.

In order to annotate part of the D-Coi corpus, the PropBank guidelines have been revised on the basis of the annotation process. During the annotation process, some problems have emerged which we aim to solve in SoNaR:

1. Linguistic problems: during the annotation some phenomena have been encountered for which linguistic research does not provide a standard solution yet. They have been discarded but it now has become imperative for us to address them.

2. Interaction among levels: there are examples in which the annotation provided by the syntactic parser is not correct as in the case of a PP which was labeled as modifier by the syntactic annotation but which should be labeled as argument according to the PropBank guidelines. Furthermore, problems have been attested with respect to PP attachment, that is the syntactic representation gives sometimes correct, sometimes incorrect structures and at the semantic level it is possible to disambiguate.

One advantage of employing PropBank for the annotation of semantic roles is that it is quite suitable for `automatic` semantic role labeling. However, in the case of Dutch, there was no semantically annotated corpus available that could be used as training data. In the D-Coi project, a novel approach to rule-based tagging based on D-Coi dependency trees has been proposed [25]. More specifically, a basic mapping between nodes in a dependency graph and PropBank roles was defined. The approach is implemented in a rule-based semantic argument tagger, called

XARA (XML-based Automatic Role-labeler for Alpino-trees) [25]. The cornerstone of Xara's rule-based approach is formed by XPath expressions. A rule in XARA consist of an XPath expression that addresses a node in the dependency tree, and a target label for that node, i.e. a rule is a *(path,label)* pair. The evaluation carried out shows that XARA achieves a precision of 65,11%, a recall of 45,83% and an F-score of 53.80. In order to evaluate the labeling of XARA, the output of XARA's semantic role tagger has been compared with the manual corrected annotation of 2,395 sentences. Since rules in XARA cover only a subset of PropBank labels, recall is notably lower than precision.

After a corpus has been tagged automatically by XARA, manual annotation can be performed relatively fast, since annotators only need to correct XARA's output instead of starting annotation from scratch.

The manually corrected sentences have been used as training and test data for a Semantic Role Labeling (SRL) classification system (i.e. Tilburg Memory based learner (TiMBL)). The features employed describe the predicate (stem, voice) and the candidate argument, that is category, dependency label, POS tag, the head word and its POS tag. In comparison to experiments in earlier work, relatively few training data was available in D-Coi: the training corpus consisted of 2,395 sentences which comprise 3066 verbs, 5271 arguments and 3810 modifiers. To overcome the data sparsity problem, the classifier was trained using the leave one out (LOO) method. With this option set, every data item in turn is selected once as a test item, and the classifier is trained on all remaining items. Except for the LOO option, only the default TiMBL settings were used during training, to prevent overfitting because of data sparsity. We refer to [25], [18] for further details.

The classifier obtained a precision of 70.27% a recall of 70.59% and an F-score of 70.43. In the annotation of the SoNar treebank, our goal is to use the classifier to annotate the new set of data and improve in this way its performance.

2.3 Spatiotemporal Labeling

Applications like multidocument, and even multilingual, summarization or question answering relies to a large extent, not only on the detection of temporal and geospatial expressions in texts, but also on their disambiguation. In SoNaR, we locate eventualities on a time axis (the *when* of an eventuality) and on a contemporary, geographical map (the *where*) whenever relevant.

Within D-Coi the development of the MiniSTEx (<u>Mini</u> <u>S</u>patio<u>T</u>emporal <u>E</u>xpressions) annotation scheme for temporal and spatial annotation was started, cf. [22, 21, 23]. The version of the annotation scheme used in SoNaR[11] makes uses of the informa-

---

[11] In D-Coi, also a general spatial component was developed, which is neglected in SoNaR.

tion contained in the treebank, including the three other semantic layers.[12] Named Entity labeling, for example, is informative in order to determine whether a spatiotemporal expression is used in a literal way or as a metonym (which affects the annotation).

The annotation scheme reflects the state of the art in geospatial and temporal annotation. With respect to the latter, TimeML [20] and TIDES [8] come to mind. Geospatial annotation as such is far less widespread and standardized. Only recently a scheme for geospatial annotation came out: SpatialML [1]. However, the subtask of disambiguation is also a subject in geographic information extraction. Some approaches in this field can be found in [7, 16, 27]. But especially as far as the Netherlands and Belgium are concerned[13] the geospatial component of the MiniSTEx scheme goes into more detail, explaining (verbatim) where something is located, for example stating that *Haren* is part of the municipality of *Borgloon* which is in the province of *Limburg* etc.[14] This way reasoning is advanced as the annotated corpus should be self-contained, i.e. the user should be able to understand a text without having a full spatiotemporal database at his disposal. In the annotation process itself, however, such a database plays a central role, as it contains the common spatiotemporal knowledge of the intended (Flemish/Dutch) audience. Other spatiotemporal knowledge is expected to be contained in the texts, although data obtained during annotation will be used to further populate the database.
With respect to the temporal component, the scheme developed within D-Coi is more detailed than, for example, TimeML in that it makes explicit between which dates *Easter* may fall, or when it was *Easter* in a specific year, taking into account the country and religion under consideration.

Contrary to TimeML and SpatialML, MiniSTEx handles temporal and (geo)-spatial annotation in one go, using a similar approach. It also handles *geotemporal* expressions, i.e. expressions associated with a combination of geospatial and temporal properties (for example to express that between the First and the Second World War *Libya*, nowadays an independent country, was a province of Italy).
Another characteristic of MiniSTEx is that full advantage is taken of the fact that the origin of the texts is known as the metadata contain the date (sometimes even the time) and place of publication, and also the title of the source (newspaper etc). From the latter the `background` of the text can be determined, and thus the `intended audience` of the text can be inferred, which to a large extent deter-

---

[12]In another project, an implementation is investigated in which no other layers are available but tagging and chunking.

[13]And to a lesser extent also neighbouring countries, and other countries the intended audience of a text (cf. [22]) is expected to be familiar with.

[14]In case the value for one of these inbetween levels is unknown, especially for entities in countries the intended audience is not supposed to be very familiar with, *XX* is used.

mines how a spatiotemporal expression is interpreted, taking into account Grice's maximes which can be paraphrased as "Don't say too much and don't say too little." The most obvious interpretation of a (spatiotemporal) expression often will not be clarified by the author, whereas other interpretations will. This information is used to calculate which interpretation of *Dover* is likely to be meant (UK, USA?), or which date(s) should be associated with, for example, *Thanksgiving* (USA, Canada?) or *summer* (which hemisphere?). In case of *Dover*: which town in which country will be referred to in a specific text does not depend on its size or population [27], but on the intra- and extratextual context in which the name is used. In a Dutch or Belgian text, a plain *Dover* will refer to the city in the UK, not to the capital of Delaware in the USA, although the latter has the larger population. A last characteristic is that quite some attention is paid to (properties with respect to) tense & aspect of the language under consideration (Dutch within SoNaR).

2.4   Apparent contradictions between layers

Labels attributed in the various layers may mesh with each other. For example, an adverb of repetition (like *again* (opnieuw, weer)) will get a label ArgM-TMP to indicate its semantic role. But in general this type of ArgM-TMP does not qualify for a spatiotemporal tag. The same holds for abstract locations (like "in his speech"), which will get a label ArgM-LOC but no (geo)spatial tag. Sometimes they may also mesh with the underlying syntactic annotation. This has to do with the fact that all annotations were developed separately, so their definitions of related concepts may differ to some degree. For example, what is considered an argument/complement and what a modifier differs between syntactic analysis and semantic role labeling. Roughly, the first considers elements to be arguments in case they can not be left out without causing ungrammaticality, whereas in the PropBank approach we adopted "a semantic role is being marked as an argument, if it frequently occurs in a corpus and is specific to a particular class of verbs" [2]. Both approaches will not necessarily lead to the same result. The question is whether this is problematic. Our feeling is that it is not a problem as long as the differences are of a systematic nature and well documented.

2.5   Too much information?

The structures the users are confronted with can be very complex when all types of semantic information are shown. The idea, however, is that although all information is present at the level of the xml-structures, the user can select one of more types of information to be shown in the trees.
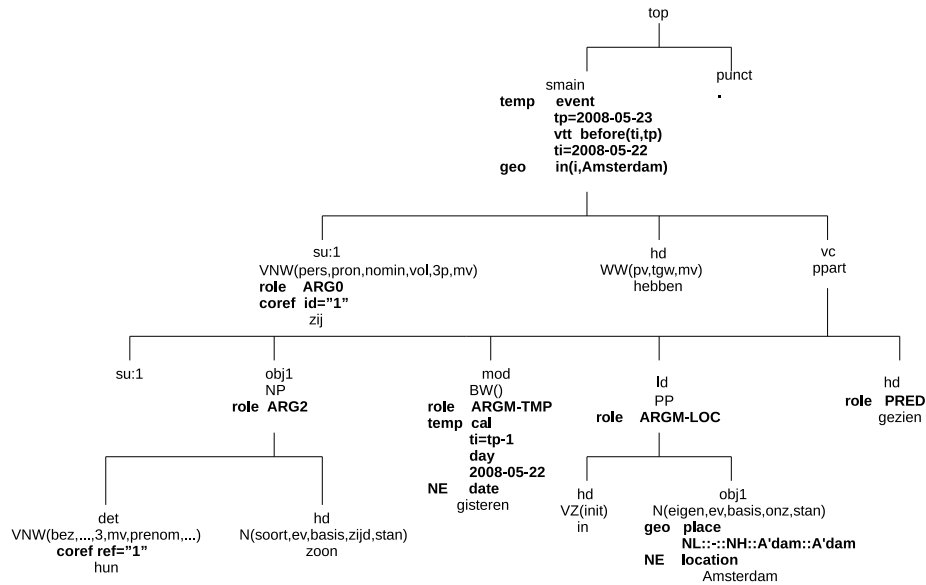
```
                                    top
                          smain                    punct
                  temp    event                      .
                          tp=2008-05-23
                          vtt  before(ti,tp)
                          ti=2008-05-22
                  geo     in(i,Amsterdam)

              su:1                        hd                vc
      VNW(pers,pron,nomin,vol,3p,mv)   WW(pv,tgw,mv)       ppart
      role   ARG0                      hebben
      coref  id="1"
              zij

   su:1      obj1              mod              ld                hd
             NP                BW()             PP           role   PRED
        role  ARG2       role   ARGM-TMP   role   ARGM-LOC         gezien
                         temp   cal
                                ti=tp-1
                                day
                                2008-05-22
                         NE     date
                                gisteren
     det           hd                     hd           obj1
VNW(bez,...,3,mv,prenom,...)  N(soort,ev,basis,zijd,stan)  VZ(init)   N(eigen,ev,basis,onz,stan)
   coref ref="1"       zoon              in      geo   place
      hun                                              NL::-::NH::A'dam::A'dam
                                                NE    location
                                                      Amsterdam
```

Figure 1: An enriched tree (simplified)

3   Conclusions

With SoNaR a whole series of manually corrected layers of annotation (from part of speech tagging over syntactic analysis to several semantic annotations) are becoming available for the same corpus of 1 million words. In se the same schemes are used in several other (STEVIN) projects, thus, especially when (part of) these data are manually corrected, enlarging the amount of data available for each type of annotation.

With respect to correction, choices were made: have everything corrected by two persons (inter-annotator agreement) or have everything corrected by one person with just parts by more persons (random consistency checks). For semantic role labeling the first option was choosen (smaller corpus), for the other layers the second option. Note that in all cases the (corrected) labels of 'previous' layers will be taken into account (errors will be reported and corrected), but that some relations are stronger than others: whereas, for example, (time-indexed) coreference labeling is essential for spatiotemporal labeling, semantic role labeling is not. Therefore, coreference is taken into account in the MiniSTEx tool, whereas the roles ArgM-LOC and ArgM-TMP are only used of in the correction phase,

References

[1] SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language, October 1 2007. MITRE Corporation.

[2] O. Babko-Malaya. *Guidelines for Propbank framers*, September 2005.

[3] S. Buchholz and A. van den Bosch. Integrating seed names and n-grams for a named entity list and classifier. In *Proceedings of LREC-2000*, pages 1215–1221, 2000.

[4] N. Chinchor. MUC-7 Named Entity Task Definition Version 3.5 (web). 1997.

[5] S. Davies, M. Poesio, F. Bruneseaux, and L. Romary. Annotating coreference in dialogues: Proposal for a scheme for MATE. http://www.hcrc.ed.ac.uk/ poesio/MATE/anno_manual.htm, 1998.

[6] F. De Meulder and W. Daelemans. Memory-based named entity recognition using unannotated data. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pages 208–211, 2003.

[7] J. Ding, L. Gravano, and N. Shivakumar. Computing Geographical Scopes of Web Resources. In *26th International Conference on Very Large Databases, VLDB 2000*, Cairo, Egypt, September 10–14 2000.

[8] L. Ferro, L.d Gerber, I. Mani, B. Sundheim, and G. Wilson. *TIDES 2005 Standard for the Annotation of Temporal Expressions*, 2005.

[9] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. Named entity recognition through classifier combination. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003*, volume 168-171, 2003.

[10] I. Hendrickx, G. Bouma, F. Coppens, W. Daelemans, V. Hoste, G. Kloosterman, A.-M. Mineur, J. Van Der Vloet, and J.-L. Verschelde. Coreference Resolution for Extracting Answers for Dutch. In *Proceedings of LREC 2008*, 2008.

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

[12] V. Hoste and W. Daelemans. Learning Dutch Coreference Resolution. In *Proceedings of CLIN 2004*, pages 133–148, 2004.

[13] V. Hoste and G de Pauw. KNACK-2002: a richly annotated corpus of Dutch written text. In *Proceedings of LREC 2006*, 2006.

[14] C.R. Johnson, C.J. Fillmore, M.R.L. Petruck, C.F. Baker, M.J. Ellsworth, J. Ruppenhofer, and E.J. Wood. *FrameNet: Theory and Practice*, 2002.

[15] P. Kingsbury, M. Palmer, and M. Marcus. Adding Semantic Annotation to the Penn TreeBank. In *Proceedings of the Human Language Technology Conference*, San Diego. California, 2002.

[16] J. Leidner. *Toponym Resolution in Text*. PhD thesis, University of Edinburgh, 2007.

[17] J. McCarthy. *A Trainable Approach to Coreference Resolution for Information Extraction*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst MA, 1996.

[18] P. Monachesi, G. Stevens, and J. Trapman. Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of LAW-07*, Prague. Czech Republic, 2007. ACL 2007 workshop.

[19] N. Oostdijk, M. Reynaert, P. Monachesi, G. Van Noord, R. Ordelman, I. Schuurman, and V. Vandeghinste. From D-Coi to SoNaR: A reference corpus for Dutch. In *Proceedings of LREC*, Marrakech, Morocco, 2008.

[20] R. Sauri, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. *TimeML Annotation Guidelines, version 1.2.1.*, 2006.

[21] I. Schuurman. Spatiotemporal Annotation on Top of an Existing Treebank. In K. De Smedt, J. Hajic, and S. Kuebler, editors, *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, pages 151–162, Bergen, Norway, 2007.

[22] I. Schuurman. Which New York, which Monday? The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions. In *Proceedings of CLIN 17*, 2007.

[23] I. Schuurman. Spatiotemporal annotation using MiniSTEx: How to deal with alternative, foreign and obsolete names? In *Proceedings of LREC 2008*, Marrakech, Morocco, 2008.

[24] W.M. Soon, H.T. Ng, and D.C.Y Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[25] G. Stevens. Automatic Role Labeling in a Dutch Corpus. Master's thesis, Utrecht University, 2006.

[26] K. van Deemter and R. Kibble. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637, 2000.

[27] R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *WWW2007*, Banff, Canada, May 8-12 2007.