KNACK-2002: a Richly Annotated Corpus of Dutch Written Text

Véronique Hoste and Guy De Pauw

CNTS - Language Technology Group
University of Antwerp
Universiteitsplein 1
veronique.hoste,guy.depauw@ua.ac.be

Abstract

In this paper, we introduce the annotated KNACK-2002 corpus of Dutch written text. The corpus features five different annotation layers, ranging from the annotation of morphological boundaries at the word level, over the annotation of part-of-speech tags and phrase chunks at the syntactic level to the annotation of named entities at the semantic level and coreferential relations at the discourse level. We believe the corpus is unique in the Dutch language area because of its richness of annotation layers, providing researchers with a useful gold standard data set for different NLP tasks in the domains of morphology, (morpho)syntax, semantics and discourse.

1. Introduction

In accordance with international tendencies, NLP researchers in Flanders and the Netherlands have gradually shifted their attention toward corpus-based methods. The recent development of the CGN corpus, a 10 million word corpus of Spoken Dutch (Oostdijk, 2000)[1] and research programs such as STEVIN[2] underline the increased importance of the data-driven paradigm for NLP in the Dutch language area. Despite offering a valuable information source for corpus-based NLP research however, the annotated CGN corpus does not offer a readily applicable information source for processing written text, since the intricacies of spoken language do not always translate well to the written domain.

In this paper, we introduce the annotated KNACK-2002 corpus of Dutch written text. Except for the annotation of coreferential relations, which was done manually, the corpus was annotated for the most part by tools trained on the basis of other annotated information sources, most importantly CGN.

About 25% of the corpus (50 texts) was manually corrected by human annotators. For the annotation layers that were semi-automatically produced on the basis of CGN, it therefore provides an interesting case study for the portability of the CGN annotation properties to other types of texts. We believe the corpus is unique in the Dutch language area because of its richness of annotation layers, providing researchers with a useful gold standard data set for different NLP tasks in the domains of morphology, (morpho)syntax, semantics and discourse. The following annotation layers are provided in the KNACK-2002 corpus:

- at the **word level**: morphological boundaries
- at the **syntactic level**: part-of-speech tags, phrase chunking information and some prosodic annotation
- at the **semantic level**: named entities
- at the **discourse level**: coreferential relations

[1] More information on this corpus can be found at http://lands.let.rug.cng/.

[2] http://taalunieversum.org/taal/technologie/stevin/

In the following section, we introduce the base material for the annotation, KNACK-2002. After the introduction of the corpus, we continue with a description of the different annotation layers. We first describe the annotation layers that have been semi-automatically produced: the annotation of morphological boundaries, part-of-speech tags, phrase chunks and named entities in Section 3. Next, we discusses the annotation of the corpus with coreferential information in Section 4. We conclude with a general overview of the corpus and detail some postprocessing work to be done in the near future.

2. KNACK-2002

KNACK-2002 is a corpus based on KNACK, a Flemish weekly news magazine with articles on national and international current affairs. KNACK covers a wide variety of topics in economical, political, scientific, cultural and social news. For the construction of the corpus, we used a selection of articles of different lengths, which all appeared in the first ten weeks of 2002. The corpus consists of 267 documents which are annotated with coreferential information. All documents are provided with the 5 previously named annotation layers.

For the creation of the first four annotation layers, i.e. the annotation of morphological boundaries, part-of-speech tags, phrase chunks and named entities, tools were trained on CGN and other information sources and an automatic annotation was provided for all documents. Lacking coreferentially annotated corpora for Dutch, the annotation of the fifth annotation layer, the coreference tag layer, was done manually.

From this large corpus of 267 documents, we made a random, but balanced selection of 50 documents covering different topics. We selected 10 documents covering internal politics, 10 documents on foreign affairs, another 10 documents on economy, 5 documents on health and health care, 5 texts covering scientific topics and finally 10 documents covering a variety of topics (such as sports, education, history and ecology). These 50 documents were completely manually verified. This manual verification was crucial since it helps us to determine the isolated error load of the related NLP tasks. A complex task such as coreference resolution, for example, depends on different

types of knowledge: morphological and lexical knowledge such as number agreement and knowledge about the type of noun phrase, syntactic information such as information about the syntactic function of anaphor and antecedent, semantic knowledge such as information about named entities, etc. Through these dependencies, automatic coreference resolution suffers from the error percolation from the NLP tasks it depends on. A manually verified corpus helps us to determine the errors specific to the task of coreference resolution.

3. Semi-automatic annotation layers

All of the annotation in the KNACK-2002 corpus, excluding coreferences, has been done semi-automatically. Existing annotation modules for morphological segmentation, part-of-speech tagging, shallow parsing and named entity recognition were used to provide a first classification for the words in the corpus. For 50 texts, these annotations were consequently verified and corrected by human annotators. This not only increases the speed with which the corpus can be annotated, it also ensures a more consistent annotation throughout the corpus.
The data was provided to the human annotators in the form of an excel file, using drop-down lists to a priori limit the possible corrections. This not only helped limit the learning curve for the human annotators, most of whom did not have an explicit background in NLP, it also allowed us to track the changes made to the data and provide relative accuracy figures for the automatic annotation modules.

3.1. Morphological segmentation

Morphological analysis is paramount to a wide number of NLP applications, including stemming and decompounding. Compounding in Dutch, for example, can occur though concatenation as in *pensioenspaarfonds* (English: *pension saving fund*) and through concatenation in combination with the infix /s/ as in *bedrijfsstructuur* (English: *company structure*) or in combination with the /e<n>/ infix as in *studentenorganisatie* (English: *student organization*) and *studentenkoepel* (English: *student umbrella organization*).
The most crucial task in morphological analysis is morpheme boundary detection, i.e. morphological segmentation. In the context of the FLaVoR project (Demuynck et al., 2003), a morphological segmentation system for Dutch was built that uses a memory-based learning classifier to predict morpheme boundaries (De Pauw et al., 2004). The system was trained on $\pm$380,000 flection forms from the morphological database of CELEX (Baayen et al., 1993). This system provided a first segmentation for all of the words in the corpus, including proper nouns and monomorphemic words.
A human annotator consequently corrected the mistakes in the smaller corpus of 50 texts. Around 8% of the words needed one or more changes to the morpheme boundaries. The most common mistake was an overeager prediction of morpheme boundaries. The KNACK-2002 corpus provides two layers of segmentation in the corpus:

- A segmentation on the orthographic realization of the word. For example for the word *ramenwasser* (English: *window cleaner*), this layer displays the segmentation as follows: *ram+en+wass+er*.
- Dutch morphology also involves quite a few orthographic alternations. We therefore also provide the canonical representation of each morpheme, i.e. reversing the orthographic changes caused by morphological process. For the word *ramenwasser* we provided the following segmentation: *raam+en+was+er* (English: *window+s+clean+er*).

3.2. Part-of-Speech Tagging

The morphosyntactic annotation of the KNACK-2002 corpus is one of the most important annotation layers. We used a number of data driven taggers trained on the CGN part-of-speech tag annotation (Van Eynde et al., 2000). Hoste (2005), however, points out that a CGN based tagger often provides awkward annotation. This is due to the fact that some of the annotation properties for spoken language do not necessarily translate well to written text. In particular, there was considerable overgeneration of the part-of-speech tag `SPEC(deeleigen)` (part of a proper noun).
We however chose to mirror the CGN part-of-speech tag annotation to the KNACK-2002 corpus. This allows us to make a useful dataset for the extensive research community currently working with CGN annotation, but also enables us to provide some insight into the portability of CGN annotation to written text.
The part-of-speech classes of the CGN corpus are rich. Apart from defining that a word is a pronoun (VNW), a verb (WW) or something else, a part-of-speech tag contains several other features of the word, as illustrated in the following sentence from the KNACK-2002 corpus.

Woensdag/N(eigen,ev,basis,zijd,stan)
waren/WW(pv,verl,mv)
gevechten/N(soort,mv,basis)
uitgebroken/WW(vd,vrij,zonder)
tussen/VZ(init)
aanhangers/N(soort,mv,basis)
van/VZ(init)
twee/TW(hoofd,prenom,stan)
lokale/ADJ(prenom,basis,met-e,stan)
rivalen/N(soort,mv,basis) ./.

English: On Wednesday, there were fights between followers of two local rivals.

We trained 3 taggers on the CGN data: MBT (Daelemans et al., 1996; Daelemans et al., 2003), TnT (Brants, 2000) and MXPOST (Ratnaparkhi, 1996). We used a stacked classifier to provide the most likely tag for the words in the corpus. The human annotators were presented with this tag, but also with the individual classifiers' decisions to speed up correction. If none of the provided tags were correct, the annotators chose the correct one from a drop-down list of possible tags.
3.4% of the words in the smaller KNACK-2002 corpus were corrected by the human annotators. They were also

asked to make minor adjustments to the annotation to facilitate the transition from spoken language to written text. This includes the aforementioned `SPEC(deeleigen)` tag which was unequivocally changed to proper noun.

3.3. Phrase Chunking

Rather than opting for a standard phrase structure approach, we have chosen phrase chunking as the syntactic representation in the KNACK-2002 corpus. Instead of building a complete tree structure, we identify phrase boundaries and phrase labels. This type of analysis can be performed using the same tools used for part-of-speech tagging, as it is a classification task that is performed on the word level. We identify the usual typical phrases: NP, VP, PP, AP, ADVP and a small number of special types of phrases:

- **VG:** conjunction words/word groups.
- **TSW:** interjections like "uh". Very uncommon in written text.
- **DETP:** cluster of determiners, e.g. "[DETP *all die*] *dingen*" (English: *all those things*)
- **MWU:** multi-word unit, i.e. a cluster of words that do not belong to any other type of phrase, e.g. "[MWU *min of meer*]" (English: *more or less*)
- **O:** not inside any of the identified phrases.

As shown in the following sentence, the chunks are base chunks. The NP chunks are base NPs, which contain a head, optionally preceded by premodifiers, such as determiners and adjectives. Postmodifiers such as "over het grensgebied" are not part of the noun phrase.

> Het/I-NP conflict/I-NP over/I-PP het/I-NP grensgebied/I-NP is/I-VP zo/I-AP oud/I-AP als/I-VG India/I-NP en/I-VG Pakistan/I-NP .
>
> English: Since the beginning of the eighties, the situation became even more dangerous: both India and Pakistan had nuclear weapons.

Using phrase-chunking information converted from the syntactically annotated files of CGN (Canisius and van den Bosch, 2004), the previously mentioned tagging systems were used in a similar vein to provide shallow parsing tags for the words in the corpus. The human annotators were then expected to correct the errors. Reviewing of the data shows about 7% of the words needed to be corrected.

3.4. Named Entities

Named entity annotations describe whether a particular word identifies a person, an organization, a location or another type of named entity. To preprocess the corpus, we used an adjusted version of the named-entity recognizer described in De Meulder et al. (2002). This recognizer combines combines gazetteers, handcrafted rules, and machine learning on the basis of seed material. A human annotator corrected this annotation layer for the KNACK-2002 corpus and also provided time reference annotation for this corpus. The resulting annotation looks as follows:

> Premier/I-PER
> Verhofstadt/I-PER
> vertrok/O
> gisteren/I-TIME
> naar/O
> New/I-LOC
> York/I-LOC
> om/O
> de/O
> UN/I-ORG
> toe/O
> te/O
> spreken/O
>
> English: Yesterday, prime minister Verhofstadt left for New York to address the UN.

3.5. Prosody

A small portion (21,000 words) of the KNACK-2002 corpus was annotated with prosodic information. For each of the words, we indicate whether or not a prosodic boundary (break, **B**) followed the word and/or whether the word received prosodic accent (**A**). Preprocessing of the data was done with an adjusted version of the prosodic annotation system developed in the PROSIT project (Marsi et al., 2003). This system automatically provided for each of the words one of four tags: (break, accent, break+accent, neither).

Since especially for written text, there are numerous correct possible annotations for each sentence, we had two annotators work in parallel on the same text. There was a significantly higher amount of annotator consensus for this corpus than was reported for prosodic annotation of the CGN corpus[3]. The result annotation is illustrated in the following example:

> De **stoet/A** van doorluchtige **popiconen/B** , **vastgeklonken/A** aan het cliché van de roerige sixties...
>
> English: the **parade** of vacuous **pop icons**, **tied** to the cliché of the swinging sixties...

4. Coreferential relations

The annotation of corpora with coreferential[4] information is useful from both a linguistic and computational point of view. From a linguistic perspective, coreferentially annotated corpora provide insight in the frequency of different types of coreferences, the type of relations between

[3] Interestingly, this may be due to the fact that, incidentally, the two annotators were identical twins.

[4] The discussion whether or not a given referring link between two constituents can be qualified as coreferential, anaphoric or both is beyond the scope of this paper.

them, etc. From a computational perspective, these corpora can be used for both the development and evaluation of automatically trained systems and for the evaluation of knowledge-based coreference resolution systems. For Dutch, no large-coverage training corpus was available that encoded the coreferential relations between noun phrases. The existing corpora for Dutch (op den Akker et al., 2002; Bouma, 2003) only contain anaphoric relations for pronouns and are rather small. The annotated corpus of op den Akker et al. (2002), for example, consists of a small number of texts from different types (newspaper articles, magazine articles and fragments of books) and only contains 801 annotated pronouns. Another small corpus for Dutch was annotated by Bouma (2003). It is based on the Volkskrant newspaper and contains anaphoric relations for 222 pronouns.

4.1. Annotated relations

For the annotation of the data, two main decisions had to be taken.

It first had to be decided between what type of constituents coreference relations would be annotated. We limited the annotation to pronouns and noun phrases. Lacking substantial Dutch corpora provided with coreferential information, not only for pronouns, but also for named entities, definite and indefinite noun phrases, the coreference annotation layer in the KNACK-2002 was manually annotated from scratch. In the corpus, 12,546 noun phrases are annotated with coreferential information.

A second decision to be taken was to decide on the type of coreferential relations. There are many different types of coreferential relations which can be encoded for noun phrases (see for example (McCarthy, 1996)): identity relations, type/token relations, part-whole/ element-set relations, nominal ellipsis, etc. When deciding on the type of relations to be annotated, it has to be taken into account that the annotation of coreferential relations is complex and can lead to disagreement among the annotators. In order to reduce the number of annotation errors, many annotation schemes (e.g. MUC-6 and MUC-7) aim at reducing the complexity of the relations to be annotated. This was also the approach we took.

For the development of the annotation scheme (Hoste, 2005), we took the MUC-7 (Hirschman and Chinchor, 1998) manual and the manual from (Davies et al., 1998) as source. We also took into account the critical remarks on these guidelines from (Kibble, 2000) and (van Deemter and Kibble, 2000). We used MITRE's "alembic Workbench"[5] as the annotation environment. As the MUC-6 (MUC-6, 1995) and MUC-7 (MUC-7, 1998) corpora, the KNACK-2002 corpus was marked with coreferential chain information. The coreferential chains are sequences of noun phrases referring to each other (idea of transitivity).

Whereas the MUC annotation scheme only describes one type of relation, namely the identity relation, we also marked other types of coreference relations, namely bound, identity of sense and a limited number of modality relations. We will now briefly discuss these types of relations. For a description of other distinctive features (e.g. the annotation of time-dependent identities and appositions) of the KNACK annotation guidelines, we refer to (Hoste, 2005).

- In case of an **identity relation**, the anaphor refers to the same referent as its antecedent, as in the following sentence as in

  > **Xavier Malisse** heeft zich geplaatst voor de kwartfinales in Wimbledon. **De Vlaamse tennisser** zal spelen tegen een onbekende tegenstander.
  >
  > English: **Xavier Malisse** has qualified for the semi-finals in Wimbledon. **The Flemish tennis player** will play against an unknown opponent.

  In the previous example, there is an identity relation between *"Xavier Malisse"* and *"De Vlaamse tennisser"*.

- As in the MUC annotations, we also marked a coreference relation between **a bound anaphor** and the NP which binds it, as in the follwing example:

  > **Geen enkele Argentijn** kan meer dan 1100 euro per maand van **zijn** rekening halen.
  >
  > English: **No Argentine** can withdraw more than 1100 euro per month from **his** bank account.

  Taking into account the critical remarks from (Kibble, 2000) and (van Deemter and Kibble, 2000) that we cannot consider this type of relation as an identity relation, we defined a new type of relation (as also proposed by Davies et al. 1998): "BOUND".

- Frequently, anaphors (such as the so-called "paycheck pronouns") do not refer to the same referent as their respective antecedents, as in the example sentence below. In this example, there is no identity relation between the antecedent noun phrase *"time credit contributions"* and the referring noun phrase *"those of the federal government"*. In order to capture this type of relationships, we follow the definition of (Hirst, 1981) and distinguish between **identity of sense anaphora (ISA)** and **identity of reference anaphora (IRA)**. An IRA (in the MUC and in our annotation scheme: "IDENT") is an anaphor which denotes the same identity as its antecedent. An ISA anaphor does not denote the same entity as its antecedent, but one of a similar description.

  > Enkele dagen eerder immers had de Waalse regering de voet op de institutionele rem gezet om een einde te maken aan **de tijdskredietpremies** die de Vlaamse regering betaalt bovenop **die van de federale overheid**.

[5] http://www.mitre.org/tech/alembic-workbench/

English: A couple of days before, the Walloon government put a break on further splitting up the institutions in order to end **the so-called "time credit" contributions** which are paid by the Flemish government on top of **those of the federal government**.

- We did also record coreference when the coreferential relation between two noun phrases is marked as possible rather than effective, as in the example below. This type of coreferential relations is marked with the "MOD" attribute.

  Schiphol, **tot op heden de meest waarschijnlijke overnemer van BIAC**, heeft zijn bod ingetrokken.

  English: **Schiphol**, **until now the most likely candidate for taking over BIAC**, has withdrawn its bid.

For a more detailed description of the annotation guidelines, we refer to (Hoste, 2005).

4.2. Consensus annotation

All 267 texts were annotated by two annotators from a pool of five native speakers with a background in linguistics. Instead of working with different possible annotations, the annotators verified all annotations together is order to reach one single consesus annotation. In case of no agreement the relation was not marked. This decision was based of the observations of (Hirschman et al., 1997) on the MUC-6 data that more than half (56%) of the errors were missing annotations and that 28% of the errors represented easy errors such as the failure to mark headlines.
The KNACK-2002 corpus annotated with coreferential information compares favorably to the existing coreferentially annotated corpora for English, which are mostly small (MUC-6 and MUC-7 contain annotations for 60 and 50 documents, respectively) and for which there is still need for much more annotation efforts.

5. Future Work and Summary

In the near future, a relation finding module will be trained on the CGN data and applied to the KNACK-2002 corpus. This will provide relational information for instance between a subject NP and its verb. This annotation layer will round up the annotation of the 125,000 word subset of the KNACK-2002 corpus. The other part of the corpus, currently only manually annotated for coreferences, will be automatically annotated again, this time using new information sources. We will also conduct experiments using annotation modules specifically trained on the manually verified core corpus of KNACK-2002, described in Section 3.. With respect to the annotation of coreferential relations, two extensions will be made to the existing annotations, viz. the annotation of the pleonastic and anaphoric use of the pronoun "het" (English: "it") and the markup of the linguistic gender of antecedents[6].
In this paper, we presented the KNACK-2002 corpus consisting of 267 documents provided with 5 annotation layers, being the annotation of morphological boundaries, part-of-speech tags, phrase chunks, named entities and coreference relations. For about 125,000 words of the corpus, a manual correction was done. The five annotation layers provide an interesting collection of gold standard data for various NLP tasks. It helps researchers to pinpoint the interdependence of information sources in classification tasks and for instance investigate whether named-entity recognition can aid in the prediction of morpheme boundaries[7]. We believe that the KNACK-2002 corpus facilitates this type of research for Dutch, while providing an interesting alternative to large-coverage corpora in its focus on a wide array of annotation layers, ranging from morphology up to the level of discourse.

Availability

The corpus will be made available on-line for academic use only and after registration at the following address: http://www.cnts.ua.ac.be/cnts/knack2002

Acknowledgments

The annotation described in this paper was funded by the NWO-FWO PROSIT project and the OCAPI BOF UA-2005 project. We would also like to thank our annotators: Anouk Holsters, Linn Melens, Steven Rossiers, Christophe Van Puymbroeck, Philip Van Puymbroeck.

6. References

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

G. Bouma. 2003. Doing dutch pronouns automatically in optimality theory. In *Proceedings of the EACL 2003 Workshop on The Computational Treatment of Anaphora*.

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

S Canisius and A. van den Bosch. 2004. A memory-based shallow parser for spoken dutch. In *Selected papers from the Thirteenth Computational Linguistics in the Netherlands Meeting*, pages 31–45.

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

W. Daelemans, J. Zavrel, A. van den Bosch, and K. van der Sloot. 2003. Memory based tagger, version 2.0, reference guide. Technical Report ILK Technical Report - ILK 03-13, Tilburg University.

[6]The Dutch male and female pronouns "hij", "hem", "zijn" and "haar" cannot only refer to living creatures but also to objects, organizations, etc.

[7]Named entities are typically monomorphemic.

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

Fien De Meulder, Walter Daelemans, and Véronique Hoste. 2002. A named entity recognition system for dutch. In *Computational Linguistics in the Netherlands 2001*, Selected Papers from the Twelfth CLIN Meeting, Amsterdam, pages 77–88, New York. Rodopi.

G. De Pauw, T. Laureys, W. Daelemans, and H. Van hamme. 2004. A comparison of two different approaches to morphological analysis of dutch. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 62–69, Barcelona, Spain.

K. Demuynck, T. Laureys, D. Van Compernolle, and H. Van hamme. 2003. FLaVoR: Flexible architecture for lvcsr. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1973–1976, Geneva, Switzerland.

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

L. Hirschman, P. Robinson, J. Burger, and M. Vilain. 1997. Automating coreference: The role of annotated training data. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

G. Hirst. 1981. Anaphora in natural language understanding: A survey. In *Lecture Notes in Computer Science*, volume 119. Springer-Verlag Berlin Heidelberg New York.

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

R. Kibble. 2000. Coreference annotation: Whither? In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, pages 1281–1286.

E. Marsi, M. Reynaert, W. Daelemans A. van den Bosch, and V. Hoste. 2003. Learning to predict pitch accents and prosodic boundaries in Dutch. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 489–496, Sapporo, Japan.

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

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

MUC-7. 1998. Muc-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

N. Oostdijk. 2000. The spoken dutch corpus. overview and first evaluation. In *Proceedings of LREC-2000 (Second International Conference on Language Resources and Evaluation)*, pages Vol. II: 887–894.

H.J.A op den Akker, M. Hospers, D. Lie, E. Kroezen, and A. Nijholt. 2002. A rule-based reference resolution method for dutch discourse. In *Proceedings 2002 Symposium on Reference Resolution in Natural Language Processing*, pages 59–66.

A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. of the Conference on Empirical Methods in Natural Language Processing, May 17-18, 1996, University of Pennsylvania*.

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

Frank Van Eynde, Jakub Zavrel, and Walter Daelemans. 2000. Part of speech tagging and lemmatisation for the spoken dutch corpus. In *Proceedings of LREC-2000*. Athens, Greece.