Northern European Journal of Language Technology, 201 , Vol. , Article, pp –19
DOI 10.3384/nejlt.2000-1533.1 1

SUC-CORE: A Balanced Corpus Annotated with Noun Phrase Coreference

Kristina Nilsson Björkenstam

Stockholm University
Computational Linguistics, Department of Linguistics
kristina.nilsson@ling.su.se

Abstract

This paper describes SUC-CORE, a subset of the Stockholm Umeå Corpus and the Swedish Treebank annotated with noun phrase coreference. While most coreference annotated corpora consist of texts of similar types within related domains, SUC-CORE consists of both informative and imaginative prose and covers a wide range of literary genres and domains. This allows for exploration of coreference across different text types, but it also means that there are limited amounts of data within each type. Future work on coreference resolution for Swedish should include making more annotated data available for the research community.

1 Introduction

One of the core Natural Language Processing tasks is *coreference resolution*, the task of identifying all expressions in a text that have the same referent in the (real or hypothetical) world. If we are able to build robust systems for coreference resolution, many existing applications can benefit from the additional information provided, e.g., within information access, natural language interaction, and machine translation (Grishman, 2003; Morton, 2005; Watson et al., 2003; Androutsopoulos and Aretoulaki, 2003). However, coreference resolution is generally considered a difficult NLP task in that it requires a combination of different kinds of linguistic knowledge, discourse processing, and inference procedures (see e.g., (Mitkov, 2002)). The task of coreference resolution is related to that of *anaphora resolution*, where the goal is to find and interpret words or phrases called *anaphors* that are pointing back to a previously mentioned expression in the discourse, called the *antecedent*. The anaphor and the antecedent are *coreferent* when they have the same referent in the real or hypothetical world.

In this paper, we are concerned with the annotation of noun phrase (NP) coreference. A distinction can be made between *referential NPs*, e.g., proper names and definite descriptions, and *anaphoric NPs*, e.g., pronouns. Referential NPs can refer independently of the linguistic context but may also be dependent on a preceding NP for interpretation, while anaphoric NPs typically require a linguistic antecedent in order to be correctly interpreted. In the context of coreference resolution, we want to find coreference links between NPs, some of which may be

categorized as anaphoric and some as referential. In (1), the referential NP *partiet* is coreferent with *hans Zanuparti* (lit. "his Zanu-party") and the anaphoric third person pronoun *hans* ("his") is coreferent with the referential NP *Robert Mugabe*.

(1) aa05:04-05[1] En storseger i presidentvalet, nästan rent svep i parlamentsvalet - officiellt jublade *Robert Mugabe* och ***hans* Zanuparti** på söndagen. Men bakom **partiets** stängda dörrar ...
'A big win in the presidential election, nearly a clean sweep in the parliamentary election - officially, *Robert Mugabe* and ***his* Zanu party** cheered on Sunday. But behind closed doors, **the party** ...'[2]

While anaphora and coreference can coincide (as in the case of *Robert Mugabe* and *hans*, above), not all cases of coreference are anaphoric, and not all cases of anaphora are coreferent, e.g., in (2) where the anaphoric pronoun *den* ("that") refers to *utveckligen av nationalinkomst* (lit. "the development of national income"). Here, the anaphoric relation is not identity-of-reference, but identity-of-sense:

(2) aa05:83 Kommissionen har färdigställt siffror som jämför *utvecklingen av Litauens nationalinkomst* ... från 1960-89 med *den* i Sverige och Finland.
'The commission has readied numbers that compare *the development of Lithuania's national income* ... from 1960-89 with *that* of Sweden and Finland.'

In this paper, we use the term *subsequent mention* to mean an NP that is coreferent with one (or more) of the preceding NPs, and the term *previous mention* to mean each coreferent NP preceding the anaphor. The *initial mention* is the NP that introduces a referent within the discourse.

Early approaches to automatic coreference resolution were knowledge-based (see e.g., (Hobbs, 1978), (Lappin and Leass, 1994), (Tetreault, 1999)), and while there is continued interest in rule-based solutions using linguistic knowledge (see e.g., (Mitkov, 2002), (Haghighi and Klein, 2009)) there was a marked shift to data-driven approaches during the 1990s (see e.g., (Connolly et al., 1994), (McCarthy and Lehnert, 1995)). Ng (2010) attributes this shift to the advent of statistical NLP and to the public availability of annotated corpora for development and evaluation. In particular, the coreference resolution tasks within the series of Message Understanding Conferences (MUC-6[3] and MUC-7[4]) and the Automatic Content Extraction program (ACE),[5] and recent SemEval and CoNLL shared task related to the OntoNotes project[6] have greatly influenced the field. These initiatives have resulted in annotation guidelines and tools, annotated corpora for system development, and evaluation methods.

The focus of MUC, and initially also ACE, was English coreference resolution, and thus many of the most influential approaches to coreference resolution were originally developed for English (e.g., (Soon et al., 2001)). While some constraints and preferences commonly used in English coreference resolution are language-insensitive, others are language-specific;

[1]Examples are labeled with the SUC text id, here aa05, and the sentence id, here 04 and 05.
[2]Translations are approximate.
[3]MUC 6, URL: www.cs.nyu.edu/cs/faculty/grishman/muc6.html
[4]MUC 7, URL: www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
[5]ACE, URL: www.itl.nist.gov/iad/mig/tests/ace
[6]OntoNotes, URL: www.bbn.com/ontonotes

For example, word order can be used in English pronoun resolution (Hobbs, 1978; Haghighi and Klein, 2009), but this is a less predictable factor for languages with freer word order such as Norwegian (Holen, 2007). The availability of coreference corpora in languages other than English is important to further the field.

A related issue is the availability of annotated data from different genres and domains: some types of text are more difficult to process than others, and constraints and preferences for resolution can be genre- and/or domain-insensitive, or specifically tailored for a particular genre or domain (Mitkov, 2002). Many approaches to coreference resolution are developed for, and evaluated against the news wire articles of MUC-6 and MUC-7, or the news wire and broadcast news of ACE-2. To broaden the applicability of coreference resolution, other kinds of data must be made available, and the current focus of the field is to include new languages, genres, and domains (see e.g., (Recasens et al., 2010) and (De Clercq et al., 2011)).

In this paper, we present SUC-CORE, a 20 000 word subset of the Stockholm Umeå Corpus (SUC) annotated with coreference relations between NPs.[7] This subset consists of the same documents as the evaluation set of the Swedish Treebank.[8] Thus, the coreference annotation of SUC-CORE can be combined with the part-of-speech tagging, morpho-syntactic analysis, and named entity annotation of SUC 2.0 (Källgren, 2006) or SUC 3.0 (Östling, 2012), and the syntactic analysis of the Swedish Treebank (Nivre et al., 2008). Through SUC-CORE, we offer annotated data for development and evaluation of coreference resolution for Swedish. To our knowledge, this is the only Swedish corpus with coreference annotation available for research.

The paper is structured as follows. We first give an overview of coreference annotated corpora in section 2. We describe the data selected for annotation from SUC in section 3, and the annotation process, with focus on some of the linguistic issues in coreference annotation, in section 4. In section 5, distributional statistics are presented. Finally, we briefly discuss positive and negative effects of adding coreference annotation to an existing, balanced corpus, and suggest a possible path forward.

2 Corpora annotated with coreference

Two of the most widely used data sets to date for machine learning experiments on coreference resolution are the Message Understanding Conference (MUC-6 and MUC-7) coreference data sets. The purpose of the MUC initiatives was to create data for Information Extraction development and evaluation, and to that end, data was annotated with different categories of names for the MUC Named Entity Recognition (NER) task (Chinchor, 1997), and coreference relations between NPs in the MUC Coreference Resolution task (Hirschman and Chinchor, 1997).

The MUC-6 and MUC-7 coreference annotated data, which is available through the Linguistic Data Consortium (LDC),[9] have been widely used for both supervised and unsupervised learning experiments (Cardie and Wagstaff, 1999; Soon et al., 2001; Ng and Cardie, 2002b,a; Yang et al., 2003; Hoste, 2005). The MUC initiative also resulted in the MUC-score for evaluation of equivalence classes (Vilain et al., 1995).

The MUC coreference annotation scheme focuses on NPs (called *markables*) that refer to the same entity. This is called the identity (IDENT) relation, and covers not only anaphoric and

[7] An earlier version of SUC-CORE is described in Björkenstam and Byström (2012).

[8] STB, URL: stp.ling.uu.se/∼nivre/swedish_treebank

[9] LDC, URL: projects.ldc.upenn.edu

referential NPs but also quantified NPs, predicative NPs, bound anaphors, and function-type expressions (Hirschman et al., 1997).

Van Deemter and Kibble (1999, 2000) argue that the coreference relation as defined in the MUC coreference task definition is too extended since it covers not only coreference relations between referring NPs, but also other anaphoric relations and non-referring NPs; They suggest to restrict the coreference task to the identity relation proper, even though this means less input to e.g., an IE system. Similarly, Borthen (2004b) argues that initiatives such as MUC that aim for efficiency by avoiding loss of information by e.g., marking all predicative NPs as coreferential with their subjects in positive sentences, lead to a representation of the reference phenomenon that is not linguistically plausible, and that will not generate an optimal result if used for machine learning.

The discussion of the MUC coreference scheme, and the usefulness of such data for machine learning, lead to several adaptations of the scheme for languages other than English (see e.g., (Hartrumpf, 2001) for German, (Hoste, 2005) for Dutch), with modifications in accordance with (Van Deemter and Kibble, 2000). Within the COREA project, an adapted version of the MUC scheme was used to annotate a corpus of Dutch and Flemish which covers a number of different genres (news text, speech transcripts, medical encyclopedia entries). The annotation includes coreference, as well as bridging, predicative, and bound relations between NPs (Hendrickx et al., 2008). The MUC discussion also lead to annotation schemes that focus on the linguistic aspects of the reference resolution task, e.g., the MATE scheme for anaphoric annotation (Poesio, 2004) and the BREDT scheme for Norwegian fiction (Borthen, 2004b,a); The latter scheme includes annotation of coreference, metonymy, bound anaphora, subset and superset relations, some types of bridging anaphora, and identity of sense. The BREDT scheme was used to annotate part of the Oslo Corpus of Tagged Norwegian Texts (Nøklestad, 2009), and a Swedish corpus of news text (Nilsson, 2010).

The Automatic Content Extraction (ACE) program, which began in 1999, is a descendant of the MUC project both in terms of motivations behind the project and the issues addressed. The objective of the ACE program is to develop information extraction technology of entities, relations, and events (Doddington et al., 2004). The overall ACE annotation task is to mark up entities (cf. discourse referents) belonging to a select set of entity types, relations between entities, and events these entities are involved in. The surface realizations of the entities are called *mentions* (cf. markables in MUC), and the annotation scheme differentiates between e.g., specific and generic referential mentions, and appositive or predicative attributive mentions (ACE, 2008). In the ACE entity detection and tracking task, all mentions of an entity are to be found and partitioned into equivalence classes (Doddington et al., 2004).

The ACE program has resulted in data for entity detection and tracking, relation detection and characterization, and event extraction (including cross-document and cross-language) in English, Chinese, and Arabic. Within the project, data-specific evaluation metrics have also been developed (Doddington et al., 2004; Strassel et al., 2008). The annotation guidelines, corpora, and other resources in support of the ACE program are available through LDC. The ACE data has been used in a number of experiments on coreference resolution, both data-driven and knowledge-based (see e.g., (Luo et al., 2004), (Chen and Hacioglu, 2006), (Ng, 2007), and (Haghighi and Klein, 2009)).

In recent years, new coreference corpora have been created within the OntoNotes project, comprising various genres of text (news text, weblogs, telephone conversations, broadcast news and talk show transcripts) in English, Chinese, and Arabic, with structural information on syn-

tax and predicate argument structure, and shallow semantics in terms of word senses linked to an ontology and coreference (Hovy et al., 2006).

While most coreference corpora consist of written news text, parts (consisting of broadcast news, talk show transcripts, and telephone conversations) of the ACE and OntoNotes corpora consist of spoken dialog; Other examples are the CHILD corpus of English task-oriented dialog (Stent and Bangalore, 2010), and the NXT-format Switchboard Corpus consisting of English telephone conversation with treebank annotation (Calhoun et al., 2010).

Besides the NXT-format Switchboard Corpora, there are a number of treebanks annotated with coreference in multiple languages, e.g., the Tübingen (Tüba D/Z) Treebank of German news text (Hinrichs et al., 2004), the NAIST Text Corpus of Japanese news text (Iida et al., 2007), and the AnCora-CO Corpus of Spanish and Catalan news text (Recasens and Marti, 2010). In the SemEval-2010 Shared Task "Coreference Resolution in Multiple Languages" (Recasens et al., 2010), some of these resources were used when the coreference task was extended to cover Catalan and Spanish (the AnCora-CO corpora), German (TüBa-D/Z), Dutch (KNACK (Hoste, 2005)), and Italian (LiveMemories (Rodriguez et al., 2010)).

Recently, the coreference resolution task has also been extended to cover more than NP coreference; For example, in the CoNLL-2011 Shared Task "Modeling Unrestricted Coreference in OntoNotes", the task was both entity and event coreference in the English OntoNotes data (Pradhan et al., 2011). Within the BioNLP field, corpora such as the GENIA corpus which consists of MEDLINE abstracts (Kim et al., 2003) have been used for the BioNLP Shared Task on Event Extraction (Kim et al., 2011).

3 Data selected from SUC

SUC is a balanced corpus, covering various text types and stylistic levels. It is modeled on the Brown Corpus (Bonelli and Sinclair, 2006) and similar sample corpora with two main categories of texts, *informative prose* and *imaginative prose*. The first category consists of e.g., news text, editorials, feature articles, and scientific papers, and the second category of different genres of fiction. SUC follows the general layout of Brown with 500 samples of text with a length of about 2,000 words each. These text samples are composed of excerpts from longer texts or a selection of short texts (Källgren, 2006). SUC has been released in three versions: SUC 1.0 (1997), SUC 2.0 (2006) and SUC 3.0 (Östling, 2012), and is available for research through Språkbanken at Gothenburg University.[10] The stand-off coreference annotation described here can be mapped to any of these versions of SUC.

The structural markup of SUC follows the Text Encoding Initiative P3 guidelines, with the basic units word, punctuation, sentence, and paragraph (Källgren, 2006). Each word in SUC is annotated with part-of-speech, morpho-syntactic information, and base form using the tagset described in (Ejerhed et al., 1992). The annotation has been manually corrected. The corpus is available in two formats, as exemplified in (3): a) SUC format, and b) Parole format. In SUC format, <w> is the SGML-tag used for words, <ana> stands for analysis, <ps> for part-of-speech, <m> for morpho-syntactic information, and <b> for base form. In Parole format, each word tag takes three attributes: the base form (lem), the linguistic analysis (msd), and the index (n).

[10]Språkbanken, URL: spraakbanken.gu.se

(3) a) <w n=24>jublade<ana><ps>VB<m>PRT AKT<b>jubla</w>

b) <w lem='jubla' msd='V@IIAS' n=24>jublade</w>.

Additionally, in SUC 2.0 there is a set of functionally interpreted structural tags, with attribute values selected by annotators, e.g, marking headlines, bylines, poetry, and abbreviations (Källgren, 2006). This set includes a tag for name expressions, which are marked with the element name ('name') and sub-classified with a value for the attribute type: 'person', 'animal', 'myth'(ogical), 'place' (location), 'inst' (organization), 'product', 'work' (of art), 'event', or 'other' (Wennstedt, 1995).

(4)
```
<w n=24>jublade<ana><ps>VB<m>PRT AKT<b>jubla</w>
<name type=person>
<w n=25>Robert<ana><ps>PM<m>NOM<b>Robert</w>
<w n=26>Mugabe<ana><ps>PM<m>NOM<b>Mugabe</w>
</name>
<w n=27>och<ana><ps>KN<b>och</w>
```

The name annotation of SUC is used for words and phrases that in the text function as proper names in a wide sense, and the annotation captures the semantic function of a name in a particular context: for example, geo-political names such as *Moskva* can occur in both the 'place' and the 'inst' categories depending on the context. Another source of inconsistency is that some expressions are annotated as names only in those cases where the expression is capitalized. In the SUC manual, the latter case is exemplified by *Utrikesdepartementet* ("The Foreign Ministry"), which can appear both with and without capitalization, and the manual states that the typography is taken into account, "in the hope that it reflects the way the writer regards the concerned word" (Källgren, 2006, p. 38)).

SUC has been further enriched with phrase structure as part of the Swedish Treebank (Nivre et al., 2008). A select set of text samples has been manually corrected, and constitutes a gold standard for the treebank. This gold standard set is balanced as it follows the overall composition of SUC, and there are proportional amounts of informative and imaginative prose selected from the largest categories in SUC. The selection is aimed at balance also within genres, e.g., the imaginative prose category includes literature of different style and artistic value, and there are samples of informative prose that can be categorized as belonging to the humanities (history) and the natural sciences (biology). The selection also takes demographics into account in terms of the sex of the authors. Further, the samples of news text were selected with geographical and political coverage in mind, and to this end, samples from both national and regional news papers of different political leanings are included.[11]

We decided to use this set of text samples, from what is already a richly annotated resource, for our coreference corpus. This coreference annotated subset, which we refer to as SUC-CORE, thus includes both informative and imaginative text of different genres and domains (see table 1). The informative prose category consists of six files with foreign and domestic news texts and editorials from national and regional morning dailies, magazine articles on interior design, a textbook excerpt on biology, and an academic essay. The imaginative prose section includes excerpts from four novels of different genres.

[11] Sofia Gustafson-Čapková, personal communication.

Table 1: Overview of SUC-CORE: file, genre, source, no. of tokens. Files marked with (*) consist of selections of texts.

File	Genre	Source	Tokens
I: Informative prose			
aa05	Press; political reportage* (foreign)	National daily	2056
aa09	Press; political reportage* (foreign, domestic)	Regional daily	2073
ba07	Press; editorials*	Regional	2100
ea10	Skills, trades and hobbies (interior design)*	Periodical	2194
ea12	Skills, trades and hobbies (biology)	Textbook	2017
ja06	Learned and scientific writing (humanities)	Textbook	2123
Total I:			**12563**
II: Imaginative prose			
kk14	Fiction (Tunström, G. "Det sanna livet")	Novel	2067
kk44	Fiction (Thorvall, K. "När man skjuter arbetare")	Novel	2016
kl07	Crime (Nesser, H. "Det grovmaskiga nätet")	Novel	2008
kn08	Romance (Dagsås, J. "Riddaren i mina drömmar")	Novel	2004
Total II:			**8095**
I + II total:			**20658**

4 Coreference Annotation

The definition of NP coreference by Van Deemter and Kibble (1999) states that assuming that the referring expressions α_1 and α_2 are occurrences of NPs, and that both have a unique reference in the context in which they occur:

Definition: α_1 and α_2 corefer if and only if Reference(α_1) = Reference(α_2)

Thus defined, the coreference relation is symmetrical and transitive, i.e., that if α_1 and α_2 are coreferential, and α_2 and α_3 are coreferential we can conclude that α_1 and α_3 are coreferential (Van Deemter and Kibble, 1999).

A *coreference chain* is a sequence formed by all NPs that corefer in a given text. Chains can stretch across sentence and paragraph boundaries, and across speaker boundaries within the same discourse. The coreference chains partition the NPs within the discourse into *equivalence classes* (Mitkov, 2002). During the annotation of SUC-CORE, our goal has been to ensure that the resulting equivalence classes consist of NPs with identical reference. Following (Van Deemter and Kibble, 1999), we restrict the annotation task to coreference relations, ruling out any considerations of bound anaphora, predicative, or bridging relations, and we further narrow the task to relations between NPs, that is, we do not handle relations between NPs and verbs, clauses, sentences, or larger stretches of discourse.

The annotation was performed by two annotators in collaboration, followed by final editing by the author. Due to the small scale of this project, we do not claim to present a general

analysis, but rather one that represents one person's interpretation of the text collection.

4.1 Entities and mentions

Following the terminology of the ACE program (Doddington et al., 2004), we define an *entity* as an object or a set of objects in the (real or hypothetical) world, and a *mention* as an expression which refers to an entity. The annotation task is restricted to three types of referring expressions:

- Name mentions (NAM): proper names and other named entities, e.g., *Robert Mugabe*. Tight appositions are included in the mention, e.g., *president* in the mention *president Mugabe*.
- Nominal mentions (NOM): NPs with a lexical noun, e.g., *partiets* ('the party'), or a nominalized adjective or a participle as head, e.g., *den gamle* ('the old+masc').
- Pronominal mentions (PRO) consist of personal pronouns, e.g., *hon* ('she'), demonstrative pronouns, e.g., *denna* ('this+uter (one)'), and reflexive pronouns, e.g., *sig* ('himself'/'herself'/'itself'). We also include possessives and genitives in this category.[12]

Mentions can be embedded in other mentions, e.g., *hans* ('his') in *hans vänner* ('his friends'):

(5) kn08:39 Efter att ha hört *greven* och ***hans*** vänner tala med varandra ...
'After having heard *the count* and ***his*** friends talk to each other ...'

If a definite NP functions as a tight apposition to the head of a mention, the largest stretch of the mention is annotated, including titles such as *den sydkoreanske presidenten* in (6) or attributes such as *oppositionspartiet* in (7).

(6) aa05:117 *den sydkoreanske presidenten Roh Tae Woo*
'the South-Korean President Roh Tae Woo'

(7) aa05:24 *oppositionspartiet ZUM*
'the opposition party ZUM'

When annotating complex NPs, the decision whether to include or exclude e.g., a postponed preposition phrase was based on whether the phrase could be interpreted as a restrictive modifier or not, e.g., in (8) where the preposition phrase *för ekonomisk uppgörelse* ('for economic settlement') is included in the mention.

(8) aa05:40 Enligt *den litauiska kommissionen för ekonomisk uppgörelse ...*
'According to *the Lithuanian commission for economic settlement ...*'

[12]The relative pronoun *som* ('who', 'which') is excluded from annotation in the current version of SUC-CORE.

4.2 Relations between mentions

The coreference relation is defined as identity of reference between two (or more) mentions, that is, they both refer to the same entity in the real (or hypothetical) world. During annotation, the objective has been to ensure that the annotated coreference relations are symmetrical and transitive, i.e., that the subsequent mention can be substituted for the previous mention without changing the meaning of the utterance. To this end, following (Van Deemter and Kibble, 1999), we apply a 'substitution' test where we check whether there is a change in meaning if we exchange the mention m_1 for a previous mention m_2. If m_2 has been linked to other mentions in the text, the test is applied to these mentions as well, by exchanging the mention m_1 for each of these mentions. If there is no change in meaning, we add a coreference link between m_1 and m_2. Because we rank precision over recall, we do not add links in ambiguous cases.

Coreference can occur between all three types of mentions, named mentions (NAM), nominal mentions (NOM), and pronominal mentions (PRO). For example, we annotate a link from the PRO subsequent mention *sin* ('his') to the NAM previous mention *Isaac* in (9:kk14:100), from the PRO previous mention *Du* ('You') to the NAM subsequent mention *Jakov* in (9:kk14:101), and from the NAM mention *Jakov* in (9:kk14:101) to the NOM mention *sin brors* ('his brother's') in (9:kk14:100).

(9) kk14:100 *Isaac* gjorde en paus och la filten tillrätta över ***sin*** **brors** rygg.[13]
'*Isaac* paused and streightened the blanket over ***his*** **brother's** back.'
kk14:101 – *Du* fryser väl inte, ***Jakov***?[14]
'– *You*'re not cold, are you, ***Jakov***?'

Because we have restricted the task to coreference annotation, bound anaphors or bridging anaphors are excluded. Also, we do not annotate relations with attributive mentions such as *en liten vivel* ('a small weevil') in (10). While such phrases add to the description of the discourse referent (here, a generic reference to the species *Bokbladmineraren*, 'The Beech Leaf Mining Weevil'), and thus may be of interest to e.g., an IE system, they are typically not definite enough to single out a specific referent. We leave annotation of such relations for future work.

(10) ea12:55-56 *Bokbladmineraren* är **en liten vivel** ...
'*The Beech Leaf Mining Weevil* is **a small weevil** ...'

Further, we do not annotate identity of sense-relations, negated expressions, expressions of modality, or function-type expressions that take different values depending on time and place. For example, we do not add a link between the mentions referring to the number of registered voters in 1985 (2.9 million) and 1990 (4.8 million):

(11) aa05:14-16 I valet för fem år sedan ... var *antalet registrerade* 2,9 miljoner (...) Annars kan man ju inte förklara att *siffran* nu svällt till 4,8 miljoner ...
'In the election five years ago ... *the number of registered voters* was 2.9 million (...) Otherwise, one cannot explain that *the figure* has increased to 4.8 million ...'

This type of expressions (and their values) should be annotated in a consistent manner; We leave this for future work.

[13] Anaphoric relation: *sin*, *Isaac*.
[14] Cataphoric relation: *Du*, *Jakov*; Anaphoric relation *Jakov*, *sin brors*.

4.2.1 Plural NPs with coordinated or split antecedents

A plural NPs can have another plural NP as its antecedent, but it can also have a mention consisting of two (or more) coordinated NPs as its antecedent (e.g., *de båda nordiska grannländerna* and *Sverige och Finland* in (12)), or two (or more) split antecedents (e.g., *de båda supermakterna* and *Sovjetunionen* and *USA* in (13)). We annotate links between plural NPs and coordinated antecedents, but leave annotation of plural NPs with split antecedents for future work.

(12) aa05:83-84 Kommissionen har färdigställt siffror som jämför utvecklingen av Litauens nationalinkomst(...) från 1960-89 med den i *Sverige och Finland*. Siffrorna visar att nationalinkomsten i Litauen är ungefär hälften så hög som i *de båda nordiska grannländerna*.
'The Commission has readied figures that compare the development of national income in Lithuania (...) from 1960-89 with that of *Sweden and Finland*. The figures show that the national income of Lithuania is about half that of *the two Nordic neighbors*.'

(13) aa05:97-130 Michail Gorbatjov för sitt handelsavtal med *USA* ... (...) De båda delstaterna har vad *Sovjetunionen* desperat behöver: bröd och teknik. (...) Den fortsatta avspänningen mellan *de båda supermakterna* har redan avkastat konkreta resultat ...
'Michail Gorbachev will get his trade agreement with *the USA* ... (...) The two states have what *the Soviet Union* desperately needs: bread and technology. (...) The continued detente between *the two super-powers* has already yielded concrete results ...'

4.2.2 Deictic pronouns

In quoted speech, deictic first and second person pronouns become anaphoric, and can be linked to the speaker or listener, or in the case of plural pronouns, to groups of people including the speaker and/or listener. Deictic pronouns also occur outside of quotes: in both informative (news text, editorials, magazine articles) and imaginative prose, deictic pronouns outside quotes can refer to the author, the reader, or to sets of people including the author or (in the case of imaginative prose) narrator.

In fiction with a first person narrator, the first expression the narrator uses to refer to himself (e.g., *jag*, 'I', in (17)) is annotated as the initial mention of this entity, and any subsequent mention in first person of this entity is linked to the closest preceding mention in this chain.

(14) kk14:1 Det blev en vidrig resa men eftersom det värsta ännu återstår skall *jag* fatta mig kort.
'It was a horrible journey but since the worst is yet to come I will be brief.'

In informative prose, first person plural pronouns can be used as a "royal" *vi* ('we') when a person speaks on behalf of an organization of some kind (that may be explicitly mentioned in the text), or as an "editorial" *vi* which refers to a generic person (everyman) as if the writer is speaking on behalf of his/her community. In the first case, we annotate a link between the "royal" *vi* and the name of the organization, if this is explicitly mentioned in the text, for example:

(15) aa05:53-55 Mellanskillnaden på 5 miljarder rubel skulle alltså utgöra en del i den totala skulden till *Litauen*. (...)
– Det är inte pengarna *vi* är ute efter i första hand.
'The difference of 5 million rubles would constitute part of the total debt to *Lithuania*. (...)
– It's not the money *we* are after primarily.'

In the case of "editorial" *vi*, there is typically no explicit antecedent. In such cases, we link each occurrence to the closest previous editorial *vi*, or, if there is no previous mention (as in (16)), mark the mention as a single mention:

(16) ba07:116 Och dessutom är det för landets energiförsörjning och för miljön viktigt att *vi* kan odla grödor ...
'Additionally, it is important for the country's energy supply and for the environment that *we* can grow crops ...'

In the fiction text samples, the author does not always list the members of the set referred to by a plural pronoun (and in such cases the extension of the set is typically not important for the reader). In (17), the first sentence of sample kl07[15] begins with the mention *vi* ("we"), including (at least) the speaker Van Veeteren and the listener Münster, and possibly also their co-workers:

(17) kl07:1-2 – Varför i helvete visste *vi* ingenting om den här Caen?
Van Veeteren satte igång innan Münster ens hunnit stänga dörren.
'– Why the hell didn't *we* know anything about this Caen?
Van Veeteren began before Münster even closed the door.'

In such cases, all occurrences of *vi* referring to the same (fuzzy) set are linked, but we do not link the *vi*-chain to any mention of the individual entities who are members of this set (e.g., *Van Veeteren*) because of the transitivity constraint (see also section 4.2.1).

Because the members of such sets vary across sentences and discourse segments, and only rarely are explicitly stated in the text by means of named or nominal mentions, there may be coreference chains in a text consisting of only pronouns. A reader of this sample can figure out which entities these sets of pronouns refer to, but if there are no explicit coordinated named or nominal mentions of the entities this will not be reflected in the annotation. A solution to this problem would be to add a layer of set relations akin to the superset and subset relations described in the BREDT scheme (Borthen, 2004a); We leave this for future work.

4.2.3 Generic and specific *man* ('one')

Most occurrences of the pronoun *man* ('one') in both informative and imaginative prose are generic references to "everyman", possibly including the author/narrator, as in (18; informative) and (19; imaginative).

(18) ba07:52-53 Det är den spontana slutsats *man* kan dra av presentation i media ...
'That is the spontaneous conclusion *one* can draw from presentation in media ...

[15] Sample kl07 consists of chapter 18 of the crime novel "Det grovmaskiga nätet" by H. Nesser.

(19) kk44:1 *Man* kan se henne där i köket med den skinande rena korkmattan, så blank att *man* kunde spegla sig i den.
'*One* can picture her there in the kitchen with the spotless linoleum, so shiny *one* could see one's reflection in it.'

We also find occurrences of *man* with a specific reference, e.g., in (20; informative):

(20) ba07:39 *Greenpeace* trycker på att de olagliga aktionerna syftar till att väcka uppmärksamhet, inte att verkligen stoppa det *man* ogillar.
'*Greenpeace* points out that the purpose of the illegal actions is to draw attention, not to really stop what *one* dislikes.'

We annotate cases like (20) where there is a clear referent for *man*, but not cases of ambiguous reference to a set of people including the author/narrator such as (18) and (19).

4.2.4 Generic and specific NPs

In one of the text samples, the biology text (ea12), there is a large number of generic NPs that refer to species of insects and plants, e.g.,:

(21) ea12:55-56 *Bokbladmineraren* är en liten vivel med förtjockade baklår som gör att *den* kan hoppa.
'*The Beech Leaf Mining Weevil* is a small weevil with enlarged back thighs that allow *it* to jump.'

This text is an example of a specific type of informative prose, with coreference relations between generic mentions. We annotate coreference relations on both a specific and a generic level, and we have taken care not to conflate coreference chains of generic and specific NPs.

4.2.5 Metonymy

Metonymy is when the the name of an attribute of an entity is used instead of the name of the entity itself. Through metonymy, a set of associations is transfered that may be important to the interpretation of the utterance. Following Recasens and Marti (2010), we argue that NPs with different semantic references can pragmatically corefer within a discourse through metonymy, and that this permits the annotation of coreference links in such cases. Frequent examples are the use of the name of a country, the capital of a country, or the building that is the seat of government to mean the government of that country, or the use of a name of a city to refer to a sports team. Such rhetorical devices may be used interchangeably while referring to the same entity in a discourse, e.g., the mentions *Moskva* and *Sovjetunionen* both refer to the government of the Soviet Union in the context of (22):

(22) aa05:39-42 För en knapp månad sedan meddelade *Moskva* litauerna att deras självständighet skulle kosta dem 21 miljarder rubel. (...) Med hjälp av bland annat arkiv och muntliga källor har [kommissionen] tagit reda på hur mycket *Sovjetunionen* beslagtagit i landet sedan annekteringen 1940.
'For little over a month ago, *Moscow* notified the Lithuanians that their independence would cost them 21 billion rubles. (...) With the help of among other things archives and oral sources, [the commission] has found out how much *the Soviet Union* has confiscated in the country since the annexation in 1940.'

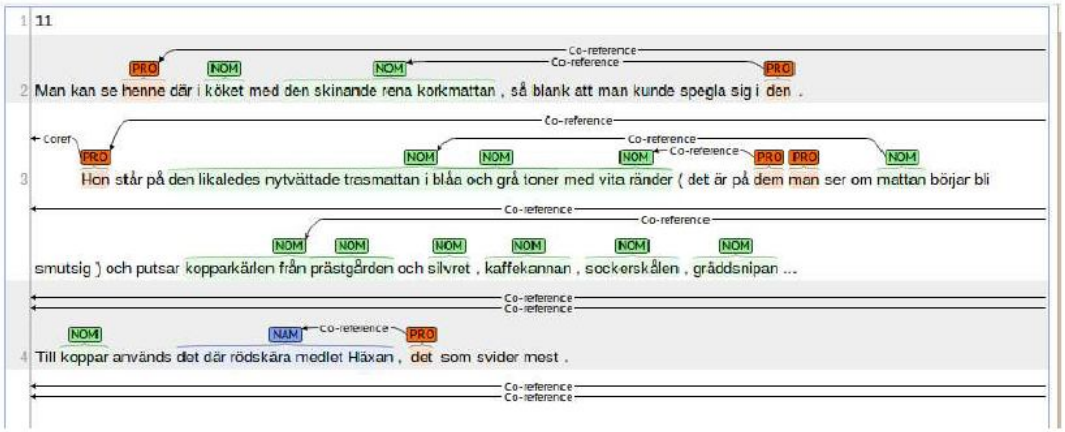

Figure 1: Screenshot of the first rows of sample kk14 in the annotation and visualization tool `brat`. Mentions of type PRO in red, NOM in green, and NAM in blue; Coreference relations are marked by directed arcs.

Mentions such as *Moskva* and *Sovjetunionen* that, in a particular text (and context), corefer on a pragmatic level through metonymy are annotated as coreference.

4.3 The annotation process

Contrary to the previously mentioned MUC, ACE, and OntoNote projects, the annotation project described here is small, both in terms of funds and, consequently, corpus size. Our corpus was created by two annotators: the author, and a student who interned in our research group as part of the course requirements of the final term of the Bachelor Program in Linguistics at Stockholm University. Because there were only two annotators available for a limited period of time, we needed a working mode that would facilitate efficient and accurate annotation.

Motivated by a study on annotation quality by Hirschman et al. (1997), who show that coreference annotation quality is improved by adopting a two-step annotation process where the first step consists of marking up all candidate mentions and the second of interpreting relations between such mentions, we decided to follow this two-step process. Thus, the annotation process was divided into a first pass in which all mentions were marked as a text span and categorized as either NAM, NOM, or PRO, and a second pass where coreference links were added between mentions with identical reference. The linking is performed by adding each subsequent (preceding) mention to the closest preceding (subsequent) coreferent mention.

The corpus was annotated by the annotators in collaboration. The annotation of mentions and relations between NPs are based on the annotators' interpretation of the textual and contextual clues, and world knowledge. Difficult cases were tried by applying the substitution test described in section 4.2. Following discussion of difficult and ambiguous cases, the author revised all annotations during a final editing pass.

4.3.1 The annotation tool

Because the selected data is already richly annotated in different formats as part of SUC and the Swedish Treebank, we decided to use stand-off annotation that may be mapped to any of these corpora.

In a pilot study, we used the ACE tool, which is available from LDC.[16] This tool has been developed to meet the demands of the ACE project, and it restricts and guides the annotation process by requiring the user to decide on one task before moving on to the next. We found that this work process was not suited for our purposes.

Instead, we decided to use the web-based annotation and visualization tool Brat (Stenetorp et al., 2012). In this tool, the underlying annotations, connected to offsets in the original text documents, are visually represented: mentions are marked with mention-type specific color highlights, and relations are marked by directed arcs between mentions. The client-server architecture of this tool allow multiple annotators to collaborate on the same documents.

In the Brat interface, a mention is annotated by selecting a span of text with the mouse, and choosing the appropriate mention type (NAM, NOM, PRO) from a dialog. Relations are annotated by dragging the mouse from one mention to the other, and choosing a pre-defined relation type from a dialog box. Existing annotations (both mentions and relations) are edited by double-clicking on a mention or a relation, and choosing the appropriate action from a dialog box. A search tool allows for searching both individual documents and the entire collection for words (by specifying whole words, substrings, or regular expressions), mention types, and relations, with results presented as concordances linked to the text documents.[17]

4.3.2 Annotation Format

The annotation is stored in a stand-off format where each mention is marked with a mention type (NAM, NOM, PRO) and connected to a specific span of text through character offsets. In (23), the following mentions are identified: *Robert Mugabe*, *hans Zanuparti* (lit. 'his Zanu party'), *hans* ('his'), *partiets stängda dörrar* (lit. 'the party's closed doors'), *partiets* ('the party').

(23) aa05:04-05 ... officiellt jublade *Robert Mugabe* och ***hans* Zanuparti** på söndagen. Men bakom **partiets** stängda dörrar ...
... officially, *Robert Mugabe* and ***his* Zanu party** cheered on Sunday. But behind **the party's** closed doors ...'

During annotation, the mention *Robert Mugabe* with index T5 is marked as mention type NAM and connected to the span 154 to 167 in the source text:

```
...
T5  NAM  154  167  Robert Mugabe
T6  NAM  172  186  hans Zanuparti
T7  NOM  212  235  partiets stängda dörrar
T8  PRO  172  176  hans
T9  NOM  212  220  partiets
...
```

[16] ACE Annotation Toolkit, URL: projects.ldc.upenn.edu/ace/tools/2005Toolkit.html

[17] We refer to the manual on the Brat web page for further information. URL: brat.nlplab.org

Table 2: Distribution of named (NAM), nominal (NOM), and pronominal (PRO) mentions in informative and imaginative prose categorized as single, initial, and subsequent mentions depending on chain position. Results presented as raw count, row percentage, and column percentage.

Count / *Row %* / *Col. %*	**I: Informative prose**				**II: Imaginative prose**			
	Single	**Initial**	**Subseq.**	***Total***	**Single**	**Initial**	**Subseq.**	***Total***
NAM	142	109	220	471	39	36	145	220
	30.2	23.1	46.7	100.0	17.7	16.4	65.9	100.0
	6.1	27.0	23.3	12.8	4.6	20.0	11.8	9.7
NOM	2133	279	377	2789	754	113	172	1039
	76.5	10.0	13.5	100.0	72.6	10.9	16.5	100.0
	91.9	69.1	39.9	76.0	88.3	62.8	14.0	45.9
PRO	47	16	349	412	61	31	911	1003
	11.4	3.9	84.7	100.0	6.1	3.1	90.8	100.0
	2.0	4.0	36.9	11.2	7.1	17.2	74.2	44.3
Col. total	2322	404	946	3672	854	180	1228	2262
	63.2	11.0	25.8	100.0	37.7	8.0	54.3	100.0
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

The coreference annotation is listed as pairwise relations between mentions, e.g., T5 (*Robert Mugabe*) and T8 (*hans*, 'his'). This directed relation can link a mention to a previous or to a subsequent mention:

```
R1  Coref  Anaphora:T8  Antecedent:T5
R2  Coref  Anaphora:T9  Antecedent:T6
...
```

This format is similar to the BioNLP Shared Task standoff format.[18] We refer the reader to the documentation of SUC-CORE and the Brat website for further details.

SUC-CORE is distributed by the Section for Computational Linguistics at the Department of Linguistics at Stockholm University.[19] SUC 2.0, SUC 3.0, and Swedish Treebank are distributed by Språkbanken at Gothenburg University.[20]

5 Distributional statistics

The distribution of named, nominal, or pronominal mentions in informative and imaginative prose in the current version of SUC-CORE is presented in table 2. Each coreferent mention is categorized as either *initial* or *subsequent* depending on its position in the coreference chain.

[18]BioNLP, URL: conll.cemantix.org/2011/data.html

[19]DALI, SU, URL: www.ling.su.se/english/nlp/resources

[20]Språkbanken, URL: spraakbanken.gu.se

A mention that introduces a referent that is referred to only once in the text is categorized as a *single* mention.

As shown in table 2, there are differences in coreference patterns between informative and imaginative prose in SUC-CORE.

First, 63.2% of all mentions in informative prose are single mentions whereas there are 37.7% single mentions in imaginative prose. Consequently, there is a larger proportion of subsequent mentions in imaginative prose (54.3%) as compared to informative prose (25.8%), and thus the coreference chains are longer in imaginative prose (average chain length 7.8 mentions, SD 1.89) than in informative prose (average chain length 3.3, SD 0.48).[21] These results indicate that for robust coreference resolution, especially for informative prose such as news texts, recognizing single mentions is as important as recognizing coreferent mentions.

Second, there are differences in the distribution of the mention types NOM and PRO: there is a larger proportion of nominal mentions in informative prose (76.0% of all mentions) than in imaginative prose (45.9%), and a smaller proportion of pronominal mentions in informative prose (11.2%) than in imaginative prose (44.3%). The conclusion we draw from this small study is that we need such diverse data in order to build robust and portable coreference resolution systems. However, further study is needed, e.g., regarding animacy, definiteness, and semantic relations between mentions, and the distribution of pronoun types in the two text types.

6 Concluding remarks

This paper describes SUC-CORE, a subset of SUC annotated with coreference relations between NPs. We decided to reuse data from SUC because it is an already richly annotated corpus which can be regarded as a standard corpus for Swedish. The stand-off annotation of SUC-CORE can be combined with the part-of-speech tagging, morpho-syntactic information, and NE annotation of SUC (Källgren, 2006), and the syntactic annotation of the Swedish Treebank (Nivre et al., 2008). Added value of reusing this data stems from the fact that SUC is freely available for research: SUC-CORE is the first publicly available Swedish corpus with coreference annotation.

To our knowledge, SUC-CORE is also the first publicly available balanced corpus with coreference annotation, consisting of both informative and imaginative prose. However, there are inherent limitations due to our choice of data. SUC-CORE consists of about 20 000 tokens in total. Of the 12 000 tokens of informative prose, there is about 6 000 tokens from news text, 2 000 from magazine articles, 2 000 from a text book on biology, and 2 000 tokens from an academic essay on history. Of the 8 000 tokens of imaginative prose, each 2 000 token sample constitutes a chapter from a different novel. The small size of each subset may be a problem when this data is used as training and test data for coreference resolution.

We acknowledge these limitations, but argue that in order to build robust and portable coreference resolution systems we need diverse data. Thus, we suggest adding more, carefully selected data from SUC to SUC-CORE. In a study on cross-domain coreference resolution, De Clercq et al. (2011) found that the amount of training data is important for resolution of out-of-domain genres, and that results can be further improved by adding genre-specific texts to the training data. They also found that data with special features (e.g., scientific texts or unedited

[21] Note that chain length is calculated from text samples of about 2000 words rather than complete discourses.

blog text) and different proportions of NP types are less suited as training data as they have less generalization power. If we can identify text samples in SUC (or other available corpora) with strong generalization power, we can build a better corpus in an efficient manner, and by adding small samples of data with special features, we can further increase the coverage and usefulness of the corpus.

Acknowledgements

Thanks to Emil Byström for participating in the development and annotation of SUC-CORE, and to Sofia Gustafson-Čapková and the anonymous reviewers for valuable comments on earlier versions of this paper.

References

ACE. 2008. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities*. LDC. Version 6.6.

Androutsopoulos, Ion and Maria Aretoulaki. 2003. Natural Language Interaction. In R. Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, chap. 35, pages 629–649. Oxford University Press.

Björkenstam, Kristina Nilsson and Emil Byström. 2012. SUC-CORE: SUC 2.0 Annotated with NP Coreference. In *Proceedings of SLTC 2012. The Fourth Swedish Language Technology Conference*. Lund, Sweden.

Bonelli, E.T. and J. Sinclair. 2006. Corpora. In K. Brown, ed., *Encyclopedia of Language and Linguistics*, pages 206–220. Oxford: Elsevier, 2nd edn.

Borthen, Kaja. 2004a. Annotation scheme for BREDT. Version 1.0. Tech. rep., University of Bergen.

Borthen, Kaja. 2004b. Predicative NPs and the annotation of reference chains. In *Proceedings of Coling 2004*, pages 1175–1178. Geneva, Switzerland.

Calhoun, Sasha, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. *Language Resources and Evaluation* 44(4):387–419.

Cardie, Claire and Kiri Wagstaff. 1999. Noun Phrase Coreference as Clustering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89. ACL.

Chen, Ying and Kadri Hacioglu. 2006. Exploration of coreference resolution: The ACE entity detection and recognition task. In *Text, Speech and Dialogue*, vol. 4188/2006 of *Lecture Notes in Computer Science*. Springer Berlin/Heidelberg.

Chinchor, Nancy. 1997. MUC-7 Named Entity Task Definition (version 3.5). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Available from http://www.itl.nist.gov/ (Last checked Oct. 14, 2005.).

Connolly, Dennis, John D. Burger, and David S. Day. 1994. A Manchine Learning Approach to Anaphoric Reference. In *Proceedings of International Conference on New Methods in Language Processing*, pages 255–261.

De Clercq, Orhee, Veronique Hoste, and Iris Hendrickx. 2011. Cross-Domain Dutch Coreference Resolution. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.

Doddington, George, 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 the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The Linguistic Annotation System of the Stockholm-Umeå Corpus Project. Tech. Rep. 33, Department of General Linguistics, University of Umeå.

Grishman, Ralph. 2003. Information Extraction. In R. Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, chap. 30, pages 545–559. Oxford University Press.

Haghighi, Aria and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: ACL.

Hartrumpf, Sven. 2001. Coreference Resolution with Syntactico-Semantic Rules and Corpus Statistics. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pages 137–144. Toulouse, France.

Hendrickx, Iris, Gosse Bouma, Frederik Coppens, Walter Daelemans, Veronique Hoste, Geert Kloosterman, Anne-Marie Mineur, Joeri Van Der Vloet, and Jean-Luc Verschelde. 2008. A coreference corpus and resolution system for Dutch. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco.

Hinrichs, Erhard, Sandra Kübler, Karin Naumann, Heike Telljohann, and Julia Trushkina. 2004. Recent developments in linguistic annotations of the TüBa-D/Z Treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany.

Hirschman, Lynette and Nancy Chinchor. 1997. MUC-7 Coreference Task Definition (version 3.0). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Available from http://www.itl.nist.gov/ (Last checked Oct. 14, 2005.).

Hirschman, Lynette, Patricia Robinson, John Burger, and Marc Vilain. 1997. Automating Coreference: The Role of Annotated Training Data. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*. AAAI.

Hobbs, Jerry R. 1978. Resolving Pronoun References. *Lingua* 44:311–338. Reprinted in Readings in Natural Language Processing, B. Grosz, K. Sparck-Jones, and B. Webber, editors, pp. 339-352, Morgan Kaufmann Publishers, Los Altos, California.

Holen, Gordana Ilić. 2007. Automatic anaphora resolution for Norwegian. In *Anaphora: Analysis, Algorithms and Applications. 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007. Lagos, Portugal, March 2007. Selected papers.*, pages 151–167. Springer.

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

Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology/North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY.

Iida, Ryu, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139. ACL, Prague.

Källgren, Gunnel. 2006. Documentation of the Stockholm Umeå Corpus. In S. Gustafson-Čapková and B. Hartmann, eds., *Manual of the Stockholm Umeå Corpus version 2.0*, pages 5–85. Department of Linguistics, Stockholm University.

Kim, J-D., T. Ohta, Y. Tateisi, and J. Tsujii. 2003. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics* 19(suppl 1).

Kim, Jin-Dong, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics, Portland, Oregon, USA.

Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–561.

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*.

McCarthy, Joseph F. and Wendy G. Lehnert. 1995. Using decision trees for coreference resolution. In C. Mellish, ed., *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, pages 1050–1055.

Mitkov, Ruslan. 2002. *Anaphora Resolution*. Longman.

Morton, Thomas. 2005. *Using Semantic Relations to Improve Information Retrieval*. Ph.D. thesis, University of Pennsylvania.

Ng, Vincent. 2007. Shallow semantics for coreference resolution. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

Ng, Vincent. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*, pages 1396–1411. ACL, Uppsala, Sweden.

Ng, Vincent and Claire Cardie. 2002a. Combining Sample Selection and Error-Driven Pruning for Machine Learning of Coreference Rules. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 55–62. ACL.

Ng, Vincent and Claire Cardie. 2002b. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111. ACL, Philadelphia, PA, USA.

Nilsson, Kristina. 2010. *Hybrid Methods for Coreference Resolution in Swedish*. Ph.D. thesis, Stockholm University.

Nivre, J., B. Megyesi, S. Gustafson-Čapková, F. Salomonsson, and B. Dahlqvist. 2008. Cultivating a Swedish Treebank. In *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, pages 111–120. Acta Universitatis Upsaliensis.

Nøklestad, Anders. 2009. *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection*. Ph.D. thesis, University of Oslo.

Östling, Robert. 2012. Stagger: A modern POS tagger for Swedish. In *Proceedings of SLTC 2012. The Fourth Swedish Language Technology Conference*. Lund, Sweden.

Poesio, Massimo. 2004. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 154–162. Boston, MA, USA.

Pradhan, Sameer, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*. Portland, Oregon.

Recasens, Marta, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, 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 Evaluation*, SemEval '10.

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

Rodriguez, K.J., F. Delogu, Y. Versley, E. Stemle, and M. Poesio. 2010. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta.

Soon, Wee Meng, 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.

Stenetorp, P., S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Sessions at EACL 2012*. ACL, France.

Stent, Amanda J. and Srinivas Bangalore. 2010. Interaction between dialog structure and coreference resolution. In *Proceedings of the Spoken Language Technology Workshop (SLT), 2010*, pages 342–347.

Strassel, Stephanie, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda. 2008. Linguistic Resource and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Tetreault, Joel R. 1999. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 602–605. Maryland, USA.

Van Deemter, Kees and Rodger Kibble. 1999. What is coreference, and what should coreference annotation be? In A. Bagga, B. Baldwin, and S. Shelton, eds., *Proceedings of the ACL Workshop on Coreference and Its Applications*. ACL, Maryland.

Van Deemter, Kees and Rodger Kibble. 2000. On Coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* 26(4):615–623.

Vilain, Marc, 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)*. Columbia, Maryland: Morgan Kaufmann.

Watson, Rebecca, Juditha Preiss, and Ted Briscoe. 2003. The contribution of domain-independent robust pronominal resolution to open-domain question answering. In *Symposium on Reference Resolution and its Applications to Question Answering and Summarization*, pages 75–82.

Wennstedt, Ola. 1995. Annotering av namn i SUC-korpusen. In K. G. Ottosson, R. V. Fjeld, and A. Torp, eds., *The Nordic Languages and Modern Linguistics 9. Proceedings of the Ninth International Conference of Nordic and General Linguistics*, pages 315–324. University of Oslo, Novis forlag.

Yang, XiaoFeng, GuoDong Zhou, Jian Su, and ChewLim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of ACL 2003, Sapporo, Japan, 7-12 July 2003*, pages 176–183.