Comparing Knowledge Sources for Nominal Anaphora Resolution

Katja Markert *
University of Leeds

Malvina Nissim †
University of Edinburgh

We compare two ways of obtaining lexical knowledge for antecedent selection in other*-anaphora and definite noun phrase coreference. Specifically, we compare an algorithm that relies on links encoded in the manually created lexical hierarchy WordNet and an algorithm that mines corpora by means of shallow lexico-semantic patterns. As corpora we use the British National Corpus (BNC), as well as the Web, which has not been previously used for this task. Our results show that (a) the knowledge encoded in WordNet is often insufficient, especially for anaphor-antecedent relations that exploit subjective or context-dependent knowledge; (b) for* other*-anaphora, the Web-based method outperforms the WordNet-based method; (c) for definite NP coreference, the Web-based method yields results comparable to those obtained using WordNet over the whole dataset and outperforms the WordNet-based method on subsets of the dataset; (d) in both case studies, the BNC-based method is worse than the other methods because of data sparseness. Thus, in our studies, the Web-based method alleviated the lexical knowledge gap often encountered in anaphora resolution, and handled examples with context-dependent relations between anaphor and antecedent. Because it is inexpensive and needs no hand-modelling of lexical knowledge, it is a promising knowledge source to integrate in anaphora resolution systems.*

1 Introduction

Most work on anaphora resolution has focused on pronominal anaphora, often achieving good accuracy. Kennedy and Boguraev (1996), Mitkov (1998), and Strube, Rapp, and Mueller (2002), e.g., report an accuracy of 75.0%, 89.7% and an F-measure of 82.8% for personal pronouns, respectively. Less attention has been paid to nominal anaphors with full lexical heads, which cover a variety of phenomena, such as *coreference* (Example (1)), *bridging* ((Clark, 1975); Example (2)), and *comparative anaphora* (Examples (3-4)).[1]

(1) The death of Maxwell, the British publishing magnate whose empire collapsed in ruins of fraud, and who was *the magazine*'s publisher, gave **the periodical** a brief international fame. (BNC)

* School of Computing, University of Leeds, Woodhouse Lane, LS2 9JT Leeds, UK. Email: markert@comp.leeds.ac.uk
† School of Informatics, University of Edinburgh, 2 Buccleuch Place, EH8 9LW Edinburgh, UK. Email: mnissim@inf.ed.ac.uk

Submission received: 15th December 2003; Revised submission received: 21st November 2004; Accepted for publication: 19th March 2005

1 In all examples, the anaphor is typed in bold face and the correct antecedent in italics. The abbreviation in parenthesis at the end of each example specifies the corpus the example is from: WSJ stands for the Wall Street Journal (WSJ), Penn Treebank, release 2; BNC stands for British National Corpus (Burnard, 1995) and MUC-6 for the combined training/test set for the coreference task of the Sixth Message Understanding Conference (Hirschman and Chinchor, 1997).

(2) [...] you don't have to undo *the jacket* to get to the map – particularly important when it's blowing a hooley. There are elasticated adjustable drawcords on the hem, waist and on **the hood**. (BNC)
(3) In addition to *increasing costs* as a result of greater financial exposure for members, these measures could have **other, far-reaching repercussions**. (WSJ)
(4) The ordinance, in Moon Township, prohibits locating *a group home for the handicapped* within a mile of **another such facility**. (WSJ)

In Example (1), the definite noun phrase (NP) "the periodical" corefers with "the magazine".[2] In Example (2), the definite NP "the hood" can be felicitously used because a related entity has already been introduced by the NP "the jacket", and a part-of relation between the two entities can be established. Examples (3-4) are instances of *other-anaphora*. *Other*-anaphora are a subclass of comparative anaphora (Halliday and Hasan, 1976; Webber et al., 2003), where the anaphoric NP is introduced by a lexical modifier (such as "other", "such", and comparative adjectives) that specifies the relationship (such as set-complement, similarity and comparison) between the entities invoked by anaphor and antecedent. For *other*-anaphora, the modifiers *other* or *another* provide a set-complement to an entity already evoked in the discourse model. In Example (3), the NP "other, far-reaching repercussions" refers to a set of repercussions *excluding* increasing costs, and can be paraphrased as "other (far-reaching) repercussions than (increasing) costs". Similarly, in Example (4), the NP "another such facility" refers to a group home which is not identical to the specific (planned) group home mentioned before.

A large and diverse amount of lexical or world knowledge is usually necessary to understand anaphors with full lexical heads. For the examples above, we need the knowledge that magazines are periodicals, that hoods are parts of jackets, that costs can be or can be viewed as repercussions of an event, and that institutional homes are facilities. Therefore, many resolution systems that handle these phenomena (Vieira and Poesio, 2000; Harabagiu, Bunescu, and Maiorano, 2001; Ng and Cardie, 2002b; Modjeska, 2002; Gardent, Manuelian, and Kow, 2003, among others) rely on handcrafted resources of lexico-semantic knowledge, such as the WordNet lexical hierarchy (Fellbaum, 1998).[3] In Section 2, we summarise previous work that has given strong indications that such resources are insufficient for the entire range of full NP anaphora. Additionally, we discuss some serious methodological problems when using fixed ontologies that have been encountered by previous researchers and/or ourselves: the costs of building, maintaining and mining ontologies, domain-specific and context-dependent knowledge, different ways of encoding information and sense ambiguity.

In Section 3, we discuss an alternative to the manual construction of knowledge bases, which we call *the corpus-based approach*. Several researchers (Hearst, 1992; Berland and Charniak, 1999, among others) suggested to enhance knowledge bases via (semi)-automatic knowledge extraction from corpora, and such enhanced knowledge bases have also been used for anaphora resolution, specifically for bridging (Poesio et al., 2002; Meyer and Dale, 2002). Building on our previous work (Markert, Nissim, and Modjeska, 2003), we extend this corpus-based approach in two ways. Firstly, we suggest to use the Web for anaphora resolution instead of the smaller sized, but less noisy and more balanced corpora used previously, making available a huge additional source of knowledge.[4] Secondly, we do not induce a fixed lexical knowledge base from the Web, but

2 In this paper, we restrict the notion of definite NPs to NPs modified by the article "the".
3 These systems also use surface level features (such as string matching), recency and grammatical constraints. In this paper, we concentrate on the lexical and semantic knowledge employed.
4 There is a growing body of research that uses the Web for NLP. As we concentrate on anaphora resolution in this paper, we refer the reader to (Grefenstette, 1999; Keller and Lapata, 2003) as well as the December

use shallow lexico-syntactic patterns and their Web frequencies for anaphora resolution on the fly. This allows us to circumvent some of the above-mentioned methodological problems that occur with any fixed ontology, whether constructed manually or automatically.

The core of this paper consists of an empirical comparison of these different sources of lexical knowledge for the task of *antecedent selection* or *antecedent ranking* in anaphora resolution. We focus on two types of full NP anaphora: *other*-anaphora (Section 4) and definite NP coreference (Section 5).[5] In both case studies, we compare an algorithm that relies mainly on the frequencies of lexico-syntactic patterns in corpora (both the Web and the BNC) with an algorithm that relies mainly on a fixed ontology (WordNet 1.7.1). We specifically adress the following questions:

1. Can the shortcomings of using a fixed ontology that were stipulated by previous research on definite NPs be confirmed in our coreference study? Do they also hold for *other*-anaphora, a phenomenon less studied so far?

2. How does corpus-based knowledge acquisition compare to using manually constructed lexical hierarchies in antecedent selection? And is the use of the Web an improvement over using smaller, but manually controlled corpora?

3. In how far is the answer to the previous question dependent on the anaphoric phenomenon addressed?

In Section 6 we discuss several aspects of our findings that still need elaboration in future work. Specifically, our work is purely comparative and regards the different lexical knowledge sources in *isolation*. It remains to be seen how the results carry forward when the knowledge sources interact with other features (for example, grammatical preferences). A similar issue concerns the integration of the methods into anaphoricity determination in addition to antecedent selection. Additionally, future work should explore the contribution of different knowledge sources for yet other anaphora types.

2 The knowledge gap and other problems for lexico-semantic resources

A number of previous studies (Harabagiu, 1997; Kameyama, 1997; Vieira and Poesio, 2000; Harabagiu, Bunescu, and Maiorano, 2001; Strube, Rapp, and Mueller, 2002; Modjeska, 2002; Gardent, Manuelian, and Kow, 2003) point to the importance of lexical and world knowledge for the resolution of full NP anaphora and the lack of such knowledge in existing ontologies (Section 2.1). In addition to this knowledge gap, we summarise other, methodological, problems with the use of ontologies in anaphora resolution (Section 2.2).

2.1 The knowledge gap for nominal anaphora with full lexical heads

In the following, we discuss previous studies on the automatic resolution of coreference, bridging and comparative anaphora, concentrating on work that yields insights into the use of lexical and semantic knowledge.

2003 special issue of Computational Linguistics for an overview of the use of the Web for other NLP tasks.

5 As described above, in *other*-anaphora the entities invoked by the anaphor are a set-complement to the entity invoked by the antecedent, whereas in definite NP coreference the entities invoked by anaphor and antecedent are identical.

Coreference The prevailing current approaches to coreference resolution are evaluated on MUC-style (Hirschman and Chinchor, 1997) annotated text and treat pronominal and full NP anaphora, named entity coreference, and non-anaphoric coreferential links that can be stipulated by appositions and copula. Their performance on definite NPs is often substantially worse than on pronouns and/or named entities (Connolly, Burger, and Day, 1997; Strube, Rapp, and Mueller, 2002; Ng and Cardie, 2002b; Yang et al., 2003). For example, for a coreference resolution algorithm on German texts, Strube, Rapp, and Mueller (2002) report an F-measure of 33.9% for definite NPs that contrasts with 82.8% for personal pronouns.

Several reasons for this performance difference have been established. First, whereas pronouns are mostly anaphoric in written text, definite NPs do not *have* to be so, inducing the problem *whether* a definite NP is anaphoric in addition to determining an antecedent among a set of potential antecedents (Fraurud, 1990; Vieira and Poesio, 2000).[6] Second, the antecedents of definite NP anaphora can occur at considerable distance to the anaphor, whereas antecedents to pronominal anaphora tend to be relatively close (Preiss, Gasperin, and Briscoe, 2004; McCoy and Strube, 1999). An automatic system can therefore more easily restrict its antecedent set for pronominal anaphora.

Third, it is in general believed that pronouns are used to refer to entities in focus, whereas entities that are not in focus are referred to by definite descriptions (Hawkins, 1978; Ariel, 1990; Gundel, Hedberg, and Zacharski, 1993). This is due to the fact that the head nouns of anaphoric definite NPs provide the reader with lexico-semantic knowledge. Antecedent accessibility is, therefore, additionally restricted via semantic compatibility and does not need to rely on notions of focus or salience to the same extent as for pronouns. Given this lexical richness of common noun anaphors, many resolution algorithms for coreference have incorporated manually controlled lexical hierarchies, such as WordNet. They use, for example, a relatively coarse-grained notion of semantic compatibility between a few high-level concepts in WordNet (Soon, Ng, and Lim, 2001), or more detailed hyponymy and synonymy links between anaphor and antecedent head nouns (Vieira and Poesio, 2000; Harabagiu, Bunescu, and Maiorano, 2001; Ng and Cardie, 2002b, among others). However, several researchers have pointed out that the incorporated information is still insufficient. Harabagiu, Bunescu, and Maiorano (2001) (see also (Kameyama, 1997)) report that evaluation of previous systems showed that "more than 30% of the missed coreference links are due to the lack of semantic consistency information between the anaphoric noun and its antecedent noun". Vieira and Poesio (2000) report results on anaphoric definite NPs in the WSJ that stand in a synonymy or hyponymy relation to their antecedent (as in Example (1)). Using WordNet links to retrieve the appropriate knowledge proved insufficient as only 35.0% of synonymy relations and 56.0% of hyponymy relations needed were encoded in WordNet as direct or inherited links.[7] The semantic knowledge used might also not necessarily improve on string matching: Soon, Ng, and Lim's (2001) final, automatically derived decision tree does not incorporate their semantic compatibility feature and instead relies heavily on string matching and aliasing, thereby leaving open how much information in a lexical hierarchy can improve over string matching.

In this paper, we concentrate on this last of the three problems (insufficient lexical knowledge). We investigate whether the knowledge gap for definite NP coreference can

6 A two-stage process where the first stage identifies anaphoricity of the NP and the second the antecedent for anaphoric NPs (Uryupina, 2003; Ng, 2004) can alleviate this problem. In this paper we focus on the second stage, namely antecedent selection.

7 Whenever we refer to "hyponymy/meronymy (relations/links)" in WordNet, we include both direct and inherited links.

be overcome by using corpora as knowledge sources as well as whether the incorporation of lexical knowledge sources improves over simple head noun matching.

Comparative Anaphora Modjeska (2002) — one of the few computational studies on comparative anaphora — shows that lexico-semantic knowledge plays a larger role than grammatical salience for *other*-anaphora. In this paper, we show that the semantic knowledge provided via synonymy and hyponymy links in WordNet is insufficient for the resolution of *other*-anaphora, although the head of the antecedent is normally a synonym or hyponym of the head of the anaphor in *other*-anaphora (Section 4.4).[8]

Bridging Vieira and Poesio (2000) report that 62.0% of meronymy relations (see Example (2)) needed for bridging resolution in their corpus were not encoded in WordNet. Gardent, Manuelian, and Kow (2003) identified bridging descriptions in a French corpus, of which 187 (52%) exploited meronymic relations. Almost 80% of these were not found in WordNet. Hahn, Strube, and Markert (1996) report experiments on 109 bridging cases from German information technology reports, using a hand-crafted, domain-specific knowledge base of 449 concepts and 334 relations. They state that 42 (38.5%) links between anaphor and antecedents were missing in their knowledge base, a high proportion given the domain-specific task. In this paper, we will not adress bridging, although we will discuss extension of our work to bridging in Section 6.

2.2 Methodological problems for the use of ontologies in anaphora resolution
Over the years, several major problems have been identified with the use of ontologies for anaphora resolution. In the following we provide a summary of the different issues raised and exemplify the problems using the examples in the Introduction.

Problem 1 – Knowledge Gap As discussed above, even in large ontologies the lack of knowledge can be severe, and this problem increases for non-hyponymy relations. All examples in Section 1 are not covered by synonymy, hyponymy or meronymy links in WordNet, e.g., hoods are not encoded as parts of jackets and homes are not encoded as a hyponym of facilities. In addition, *building*, *extending*, and *maintaining* ontologies by hand is expensive.

Problem 2 – Context-dependent Relations Whereas the knowledge gap might get reduced as (semi-)automatic efforts to enrich ontologies become available (Hearst, 1992; Berland and Charniak, 1999; Poesio et al., 2002), the second problem is intrinsic to fixed context-independent ontologies: *how much* and *which* knowledge should they include? Thus, Hearst (1992) raises the issue of whether underspecified, context- or point-of-view dependent hyponymy relations (like the context-dependent link between "costs" and "repercussions" in Example (3)) should be included in a fixed ontology in addition to universally true hyponymy relations. Some other hyponymy relations that we encountered in our studies and whose inclusion into ontologies is debatable are age:(risk) factor, coffee:export, pilots:union, country:member.

Problem 3 – Information Encoding Knowledge might be encoded in many different ways in a lexical hierarchy and this can pose a problem for anaphora resolution (Humphreys et al., 1997; Poesio, Vieira, and Teufel, 1997). For example, although "magazine" and "periodical" are not linked in WordNet via synonymy/hyponymy, the gloss records "magazine" as a "periodic publication". Thus, the analysis of the gloss together with

8 From now on, we will often use the terms "anaphor" and "antecedent" instead of "head of anaphor" and "head of antecedent" if the context is non-ambiguous.

derivation of "periodical" from "periodic" might derive the desired link. However, such extensive mining of the ontology (as e.g., performed by (Harabagiu, Bunescu, and Maiorano, 2001)) can be costly. In addition, different information sources must be weighed (e.g., is a hyponymy link preferred over a gloss inclusion?) and combined (should hyponyms/hyperonyms/sisters of gloss expressions be considered recursively?). Extensive combinations also increase the risk of false positives.[9]

Problem 4 – Sense Proliferation Using all senses of anaphor and potential antecedents in the search for relations might yield a link between an incorrect antecedent candidate and the anaphor due to an inappropriate sense selection. On the other hand, considering only the most frequent sense for anaphor and antecedent (as performed in (Soon, Ng, and Lim, 2001)) might lead to wrong antecedent assignment if a minority sense is intended in the text. So, for example, the most frequent sense of "hood" in WordNet is "criminal", whereas the sense used in Example (2) is "headdress". The alternatives are either weighing senses according to different domains, or a more costly sense disambiguation procedure before anaphora resolution (Preiss, 2002).

3 The Alternative: Corpus-based Knowledge Extraction

There have been a considerable number of efforts to extract lexical relations from corpora in order to build new knowledge sources and enrich existing ones without time-consuming hand-modelling. This includes the extraction of hyponymy and synonymy relations (Hearst, 1992; Caraballo, 1999, among others) as well as meronymy (Berland and Charniak, 1999; Meyer, 2001).[10] One approach to the extraction of instances of a particular lexical relation is the use of *patterns* that express lexical relations *structurally explicitly* in a corpus (Hearst, 1992; Berland and Charniak, 1999; Caraballo, 1999; Meyer, 2001) and this is the approach we focus on here. As an example, the pattern `NP`$_1$ `and other` `NP`$_2$ usually expresses a hyponymy/similarity relation between the hyponym NP$_1$ and its hypernym NP$_2$ (Hearst, 1992) and it can therefore be postulated that two noun phrases that occur in such a pattern in a corpus should be linked in an ontology via a hyponymy link. Applications of the extracted relations to anaphora resolution are less frequent. However, Poesio et al. (2002) and Meyer and Dale (2002) have used patterns for the corpus-based acquisition of meronymy relations, which are subsequently exploited for bridging resolution.

Although automatic acquisition can help bridge the knowledge gap (Problem 1 in the previous section), the incorporation of the acquired knowledge into a fixed ontology yields other problems. Most notably, it has to be decided which knowledge should be included in ontologies, because pattern-based acquisition will also find spurious, subjective and context-dependent knowledge (see Problem 2). There is also the problem of pattern ambiguity, since patterns do not necessarily have a one-to-one correspondence to lexical relations (Meyer, 2001). Following our work in (Markert, Nissim, and Modjeska, 2003), we argue that for the task of antecedent ranking these problems can be circumvented by *not* constructing a fixed ontology at all. Instead, we use the pattern-based approach to find lexical relationships holding between anaphor and antecedent in corpora *on the fly*. For instance, in Example (3) we do not need to know whether costs are *always* repercussions (and should therefore be linked via hyponymy in an on-

9 Even without extensive mining, this risk can be high: Vieira and Poesio (2000) report a high number of false positives for one of their datasets, although they use only WordNet encoded links.

10 There is also a long history in the extraction of other lexical knowledge, which is also potentially useful for anaphora resolution, for example of selectional restrictions/preferences. In this paper we focus on the lexical relations that can hold between antecedent and anaphor head nouns.

tology) but only that they are *more likely* to be viewed as repercussions than the other antecedent candidates. We therefore adapt the pattern-based approach in the following way for antecedent selection.

Step 1 – Relation Identification We determine which lexical relation usually holds between anaphor and antecedent head nouns for a particular anaphoric phenomenon. E.g, in *other*-anaphora, a hyponymy/similarity relation between anaphor and antecedent is exploited (homes are facilities) or stipulated by the context (costs are viewed as repercussions).

Step 2 – Pattern Selection We select patterns that express this lexical relation *structurally explicitly*. For example, the pattern `NP1 and other NP2` usually expresses hyponymy/similarity relations between the hyponym NP_1 and its hypernym NP_2 (see above).

Step 3 – Pattern Instantiation If the lexical relation between anaphor and antecedent head nouns is strong, then it is likely that they also frequently cooccur in the selected explicit patterns. We extract all potential antecedents for each anaphor, and instantiate the explicit pattern for all anaphor/antecedent pairs. In Example (4) the pattern `NP1 and other NP2` can be instantiated with `ordinances and other facilities`, `Moon Township and other facilities`, `homes and other facilities`, `handicapped and other facilities`, and `miles and other facilities`.[11]

Step 4 – Antecedent Assignment The instantiation of a pattern can be searched in any corpus to determine its frequency. We follow the rationale that the most frequent of these instantiated patterns determines the most likely antecedent. Therefore, should the head noun of an antecedent candidate and the anaphor cooccur in a pattern although they do not stand in the lexical relationship considered (due to pattern ambiguity, noise in the corpus, or spurious occurrences) this need not prove a problem as long as the correct antecedent candidate cooccurs more frequently with the anaphor.

As the patterns can be elaborate, most manually controlled and linguistically processed corpora are too small to determine the pattern frequencies reliably. Therefore, the size of the corpora used in some previous approaches leads to data sparseness (Berland and Charniak, 1999) and the extraction procedure can therefore require extensive smoothing. Thus as a further extension, we suggest using the largest corpus available, the Web, in the above procedure. The instantiation for the correct antecedent `homes and other facilities` in Example (4), for instance, does not occur at all in the BNC, but yields over 1500 hits on the Web.[12] The competing instantiations (listed in Step 3) yield 0 hits in the BNC and hits lower than 20 on the Web.

In the remainder of this paper, we present two comparative case studies on coreference and *other*-anaphora that evaluate the ontology- and the corpus-based approach in general and our extensions in particular.

4 Case Study I: *Other*-Anaphora

We now describe our first case study for antecedent selection in *other*-anaphora.

4.1 Corpus Description and Annotation

We use Modjeska's (2003) annotated corpus of *other*-anaphors from the WSJ. All examples in this section are from this corpus. Modjeska restricts the notion of *other*-anaphora

11 These simplified instantiations serve as an example; for final instantiations see Section 4.5.1.
12 This search and all searches for the Web experiments in Case Study I were executed on 29.08.2003. All Web searches for Case Study II were executed 27.08.2004.

to anaphoric NPs with full lexical heads modified by "other" or "another" (Examples (3-4)), thereby excluding idiomatic non-referential uses, e.g. "on the other hand", reciprocals such as "each other", ellipsis, and *one*-anaphora. The excluded cases are either non-anaphoric or do not have a full lexical head and would therefore require a mostly non-lexical approach to resolution. Her corpus also excludes *other*-anaphors with *structurally* available antecedents: in *list-contexts* such as Example (5) the antecedent is normally given as the left conjunct of the list.

(5) [. . .] AZT can relieve *dementia* and **other symptoms** in children [. . .]

A similar case is the construction "Xs other than Ys". For a computational treatment of *other*-NPs with structural antecedents see (Bierner, 2001).

The original corpus collected and annotated by Modjeska (2003) contains 500 instances of *other*-anaphors with NP antecedents in a five-sentence window. In this study we use the 408 (81.6%) *other*-anaphors in the corpus that have NP antecedents within a *two*-sentence window (the current or previous sentence).[13] An antecedent candidate is manually annotated as correct if it is the latest mention of the entity that the anaphor provides the set-complement to. A tag lenient was used to annotate previous mentions of the same entity. In Example (6), "all other bidders" refers to all bidders excluding United Illuminating Co., whose latest mention is "it". In this paper, lenient antecedents are underlined. All other potential antecedents, e.g. "offer" in Example (6), are called *distractors*.

(6) United Illuminating Co. raised its proposed offer to one *it* valued at $2.29 billion from $2.19 billion, apparently topping **all other bidders**.

The antecedent can be a set of separately mentioned entities, like "May" and "July" in Example (7). For such *split antecedents* (Modjeska, 2003), the latest mention of *each* set member is annotated as correct, so that there can be more than one correct antecedent to an anaphor.[14]

(7) The *May* contract, which also is without restraints, ended with a gain of 0.45 cent to 14.26 cents. The *July* delivery rose its daily permissible limit of 0.50 cent a pound to 14.00 cent, while **other contract months** showed near-limit advances.

4.2 Antecedent Extraction and Preprocessing

For each anaphor, all previously occurring NPs in the two-sentence window were automatically extracted exploiting the WSJ parse trees. NPs containing a possessive NP modifier, e.g., "Spain's economy", were split into a possessor phrase, "Spain", and a possessed entity, "Spain's economy".[15] Modjeska (2003) identifies several syntactic positions that cannot serve as antecedents of *other*-anaphors. We only automatically exclude NPs preceding an appositive *other*-anaphor from the candidate antecedent set. In "Mary Elizabeth Ariail, **another social-studies teacher**", the NP "Mary Elizabeth Ariail" cannot be the antecedent of "another social-studies teacher" as the two phrases are coreferential and cannot provide a set-complement to each other.

13 We concentrated on this majority of cases to focus on the comparison of different sources of lexical knowledge without involving discourse segmentation or focus tracking. In Case Study II we expand the window size to allow equally high coverage.

14 The occurrence of split antecedents also motivated the distinction between correct and lenient antecedents in the annotation. Anaphors with split antecedents have several antecedent candidates annotated as correct. All other anaphors have only one antecedent candidate annotated as correct, with previous mentions of the same entity marked as lenient.

15 We thank Natalia Modjeska for the extraction and for making the resulting sets of candidate antecedents available to us.

The resulting set of potential NP antecedents for an anaphor ana (with a unique identifier $anaid$) is called $\mathcal{A}_{anaid}$.[16] The final number of extracted antecedents for the whole dataset is 4272, with an average of 10.5 antecedent candidates per anaphor.

After extraction, *all modification* was eliminated and only the rightmost noun of compounds was kept, as modification results in data sparseness for the corpus-based methods, and compounds are often not recorded in WordNet.

For the same reasons we automatically resolved *Named Entities* (NEs). They were classified into the ENAMEX MUC-7 categories (Chinchor, 1997) PERSON, ORGANIZATION and LOCATION, using the software ANNIE (GATE2, `http://gate.ac.uk`). We then automatically obtained more fine-grained distinctions for the NE categories LOCATION and ORGANIZATION, whenever possible. We classify LOCATIONS into COUNTRY, (US) STATE, CITY, RIVER, LAKE and OCEAN in the following way. First, small gazetteers for these subcategories were extracted from the Web. Second, if an entity marked as LOCATION by ANNIE occurred in exactly one of these gazetteers (e.g., Texas in the (US) STATE gazetteer) it received the corresponding specific label; if it occurred in none or in several of the gazetteers (e.g. Mississippi occurred in both the state and the river gazetteer) then the label was left at the LOCATION level. We further classified an ORGANIZATION entity by using its internal make-up as follows. We extracted all single-word hyponyms of the noun "organization" from WordNet and used this set *OrgSet* as the target categories for the fine-grained distinctions. If an entity was classified by ANNIE as ORGANISATION and it has an element <ORG> of *OrgSet* as its final lemmatised word (e.g "Deutsche Bank") or contains the pattern "<ORG> of" (for example "Bank of America"), it was subclassified as <ORG> (here, BANK). In cases of ambiguity, again, no subclassification was carried out. No further distinctions were developed for the category PERSON. For numeric and time entities we used regular expression matching to classify them into DAY, MONTH, YEAR as well as DOLLAR or simply NUMBER. This subclassification of the standard categories provides us with additional lexical information for antecedent selection. Thus, in Example (8), for instance, a finer grained classification of "South Carolina" into STATE provides more useful information than resolving both "South Carolina" and "Greenville County" as LOCATION only.

(8) Use of Scoring High is widespread in *South Carolina* and common in Greenville County [...]. Experts say there isn't **another state** in the country where [...]

Finally, all antecedent candidates and anaphors were lemmatised. The procedure of extraction and preprocessing results in the following antecedent sets and anaphors for Example (3) and Example (4): $\mathcal{A}_3$={*[...], addition, cost, result, exposure, member, measure*} and ana=*repercussion* and $\mathcal{A}_4$= {*[...], ordinance, Moon Township [=location], home, handicapped, mile*} and ana=*facility*.

Table 1 shows the distribution of antecedent NP types in the *other*-anaphora dataset.[17] NE resolution is clearly important as 205 of 468 (43.8%) of correct antecedents are NEs.

4.3 Evaluation Measures and Baselines

For each anaphor, each algorithm selects at most one antecedent as the correct one. If this antecedent provides the appropriate set complement to the anaphor, i.e., is marked in the goldstandard as correct or lenient, the assignment is evaluated as correct.[18] Otherwise, it is evaluated as wrong. We use the following evaluation measures: *precision* is

16 In this paper the anaphor id corresponds to the example numbers.
17 Note that there are more correct antecedents than anaphors because the data includes split antecedents.
18 This does not hold for anaphors with split antecedents where all antecedents marked as correct need to be found in order to provide the complete set-complement. Therefore, all our algorithms' assignments in these cases are evaluated as wrong, as they select at most one antecedent.

Table 1
Distribution of antecedent NP types in the *other*-anaphora dataset.

	CORRECT	LENIENT	DISTRACTORS	ALL
pronouns	49	19	329	397
named entities	205	56	806	1067
common nouns	214	104	2490	2808
total	468	179	3625	4272

the number of correct assignments divided by the number of assignments, *recall* is the number of correct assignments divided by the number of anaphors, and *F-measure* is based on equal weighting of precision and recall. In addition, we also give the *coverage* of each algorithm as the number of assignments divided by the number of anaphors. This latter measure is included to indicate how often the algorithm has any knowledge to go on, whether correct or false. For algorithms where the coverage is 100%, precision, recall and F-measure all coincide.

We developed two simple rule-based baseline algorithms. A recency-based baseline (baselineREC) always selects the antecedent candidate closest to the anaphor. The second one (baselineSTR) takes into account that the lemmatised head of an *other*-anaphor is sometimes the same as that of its antecedent, as in "the pilot's claim ... other bankruptcy claims". For each anaphor, baselineSTR string-compares its last (lemmatised) word with the last (lemmatised) word of each of its potential antecedents. If the strings match, the corresponding antecedent is chosen as the correct one. If several antecedents produce a match, the baseline chooses the most recent one among them. If no antecedent produces a match, no antecedent is assigned. We tested two variations of this baseline.[19] The algorithm baselineSTR_{v1} uses only the original antecedents for string matching, disregarding named entity resolution. A backoff version ($\text{baselineSTR}^*_{v1}$) chooses the antecedent closest to the anaphor among all antecedent candidates if string-comparison returns no match, thereby yielding a 100% coverage. The second variation baselineSTR_{v2} uses the replacements for named entities for string matching; again a backoff version ($\text{baselineSTR}^*_{v2}$) uses a recency backoff. This baseline performs slightly better as now also cases as in Example (8) ("South Carolina ... another state", in which "South Carolina" is resolved to STATE) can be resolved. The results of all baselines are summarised in Table 2. Results of the 100% coverage backoff algorithms are indicated by $precision^*$ in all tables. The set of anaphors covered by the string matching baselines baselineSTR_{v1} and baselineSTR_{v2} will be called StrSet_{v1} and StrSet_{v2}, respectively. These sets do *not* include the cases assigned by the recency backoff in $\text{baselineSTR}^*_{v1}$ and $\text{baselineSTR}^*_{v2}$.

For our WordNet and corpus-based algorithms we additionally deleted pronouns from the antecedent sets, since they are lexically not very informative and are also not encoded in WordNet. This removes 49 (10.5%) of the 468 correct antecedents (see Table 1); however, we can still resolve some of the anaphors with pronoun antecedents if they also have a lenient non-pronominal antecedent, as in Example (6). After pronoun deletion, the total number of antecedents in our dataset is 3875 for 408 anaphors, of which 419 are correct antecedents, 160 are lenient, and 3296 are distractors.

19 Different versions of the same prototype algorithm are indicated via an index of $v1$, $v2$ [...]. The general prototype algorithm is referred to without indices.

Table 2
Overview of the results for all baselines for *other*-anaphora

algorithm	coverage	precision	recall	F-measure	$precision^*$
baselineREC	1.000	0.178	0.178	0.178	0.178
baselineSTR$_{v1}$	0.282	0.686	0.194	0.304	0.333
baselineSTR$_{v2}$	0.309	0.698	0.216	0.329	0.350

4.4 Wordnet as a Knowledge Source for *Other*-anaphora Resolution

4.4.1 Descriptive Statistics As most antecedents are hyponyms or synonyms of their anaphors in *other*-anaphora, for each anaphor ana, we look up which elements of its antecedent set $\mathcal{A}_{anaid}$ are hyponyms/synonyms of ana in WordNet, considering all senses of anaphor and candidate antecedent. In Example (4), e.g., we look up whether "ordinance", "Moon Township", "home", "handicapped", and "mile" are hyponyms or synonyms of "facility" in WordNet. Similarly, in Example (9), we look up whether "Will Quinlan" [=person], "gene", and "risk" are hyponyms/synonyms of "child".

(9) *Will Quinlan* had not inherited a damaged retinoblastoma supressor gene and, therefore, faced no more risk than **other children** [...]

As proper nouns (e.g. "Will Quinlan") are often not included in WordNet, we also look up whether the NE category of an NE antecedent is a hyponym/synonym of the anaphor (e.g., whether person is a synonym/hyponym of child) *and vice versa* (e.g., whether child is a synonym/hyponym of person). This last inverted look-up is necessary as the NE category of the antecedent is often too general to preserve the normal hyponymy relationship to the anaphor. Indeed, in Example (9), it is the inverted look-up that captures the correct hyponymy relation between person and child. If the single look-up for common nouns or any of the three look-ups for proper nouns is successful we say that a *hyp/syn relation* between candidate antecedent and anaphor holds in WordNet. Note that each noun in WordNet stands in a hyp/syn relation to itself. Table 3 summarises how many correct/lenient antecedents and distractors stand in a hyp/syn relation to their anaphor in WordNet.

Table 3
Descriptive statistics for WordNet hyp/syn relations for *other*-anaphora.

	hyp/syn relation to ana	no hyp/syn relation	total
correct antecedents	180 (43.0%)	239 (57.0%)	419 (100%)
lenient antecedents	68 (42.5%)	92 (57.5%)	160 (100%)
distractors	296 (9.0%)	3000(91.0%)	3296 (100%)
all antecedents	544 (14.0%)	3331 (86.0%)	3875 (100%)

Correct/lenient antecedents stand in a hyp/syn relation to their anaphor significantly more often than distractors do ($p < 0.001$, t-test). The use of WordNet hyponymy/synonymy relations to distinguish between correct/lenient antecedents and distractors is therefore plausible. However, Table 3 also shows two limitations of relying on WordNet in resolution algorithms. Firstly, 57% of correct and lenient antecedents are *not* linked via a hyp/syn relation to their anaphor in WordNet. This will affect coverage and recall (see also Problem 1, Section 2). Examples from our dataset that are not covered

are home:facility, cost:repercussion, age:(risk) factor, pension:benefit, coffee:export, pilot(s):union, including both missing universal hyponymy links and context-stipulated ones. Secondly, the raw frequency (296) of distractors that stand in a hyp/syn relation to their anaphor is higher than the combined raw frequency for correct/lenient antecedents (248) that do so, which can affect precision. This is due to both sense proliferation (Problem 4, Section 2) and anaphors that require more than just lexical knowledge about antecedent and anaphor heads to select a correct antecedent over a distractor. In Example (10), the distractor "product" stands in a hyp/syn relationship to the anaphor "commodity" and — disregarding other factors — is a good antecedent candidate.[20]

(10) [...] the move is designed to more accurately reflect the value of products and to put *steel* on a more equal footing with **other commodities**.

4.4.2 The WordNet-based Algorithm The WordNet-based algorithm resolves each anaphor ana to a hyponym or synonym in $\mathcal{A}_{anaid}$, if possible. If several antecedent candidates are hyponyms or synonyms of ana, it uses a tiebreaker based on string match and recency. When no candidate antecedent is a hyponym or synonym of ana, string match and recency can be used as a possible backoff.[21] String comparison for tiebreaker and backoff can again use the original or the replaced antecedents, yielding two versions algoWN$_{v1}$ (original antecedents) and algoWN$_{v2}$ (replaced antecedents).

The exact procedure for the version algoWN$_{v1}$ given an anaphor ana is as follows:[22]

(i) for each antecedent a in $\mathcal{A}_{anaid}$, look up whether a hyp/syn relation between a and ana holds in WordNet; if this is the case, push a into a set $\mathcal{A}_{anaid}^{hyp/syn}$;

(ii) if $\mathcal{A}_{anaid}^{hyp/syn}$ contains exactly one element, choose this element and stop;

(iii) otherwise, if $\mathcal{A}_{anaid}^{hyp/syn}$ contains more than one element, string-compare each antecedent in $\mathcal{A}_{anaid}^{hyp/syn}$ with ana (using original antecedents only). If exactly one element of $\mathcal{A}_{anaid}^{hyp/syn}$ matches ana, select this one and stop; if several match ana, select the closest to ana within these matching antecedents and stop; if none match, select the closest to ana within $\mathcal{A}_{anaid}^{hyp/syn}$ and stop;

(iv) otherwise, if $\mathcal{A}_{anaid}^{hyp/syn}$ is empty, make no assignment and stop.

The backoff algorithm algoWN$^*_{v1}$ uses baselineSTR$^*_{v1}$ as a backoff ([iv']) if no antecedent could be assigned:

(iv') otherwise, if $\mathcal{A}_{anaid}^{hyp/syn}$ is empty, use baselineSTR$^*_{v1}$ to assign an antecedent to ana and stop;

Both algoWN$_{v1}$ and algoWN$_{v2}$ achieved the same results, namely a coverage of 65.2%, precision of 56.8% and recall of 37.0%, yielding an F-measure of 44.8%. The low coverage and recall confirm our predictions in Section 4.4.1. Using backoff algoWN$^*_{v1}$/ algoWN$^*_{v2}$ achieves a coverage of 100% and a precision/recall/F-measure of 44.4%.

4.5 Corpora as Knowledge Sources for *Other*-anaphora Resolution

In Section 3 we suggested the use of shallow lexico-semantic patterns for obtaining anaphor-antecedent relations from corpora. In our first experiment we use the Web that

20 This problem is not WordNet-specific but affects all algorithms that rely on lexical knowledge only.

21 Because each noun is a synonym of itself, anaphors in StrSet$_{v1}$/StrSet$_{v2}$ that do have a string-matching antecedent candidate will already be covered by the WordNet lookup prior to backoff in almost all cases: backoff string matching will only take effect if the anaphor/antecedent head noun is not in WordNet at all. Therefore, the described backoff will most of the time just amount to a recency backoff.

22 The algorithm algoWN$_{v2}$ follows the same procedure apart from the variation in string matching.

with its approximately 8058M pages[23] is the largest corpus available to the NLP community. In our second experiment we use the same technique on the BNC, a smaller (100M words) but virtually noise-free and balanced corpus of contemporary English.

4.5.1 Pattern Selection and Instantiation The list-context `Xs and other Ys` explicitly expresses a hyponymy/synonymy relationship with X being hyponyms/synonyms of Y (see also Example (5) and (Hearst, 1992)). This is only one of the possible structures that express hyponymy/synonymy. Others involve "such", "including" and "especially" (Hearst, 1992) or appositions and coordination. We derive our patterns from the list-context because it corresponds relatively unambigously to hyponymy/synonymy relations (in contrast to coordination, which often links sister concepts instead of a hyponym and its hyperonym as in "tigers and lions" or even completely unrelated concepts). In addition, it is quite frequent (for example, *and other* occurs more frequently on the Web than *such as* and *other than*). Future work has to explore which patterns have the highest precision and/or recall and how different patterns can be combined effectively without increasing the risk of false positives (see also Problem 3, Section 2).

Web For the Web-Algorithm (algoWeb), we use the following pattern:[24]

(W1) `(`$N_1\{sg\}$ `OR` $N_1\{pl\}$`) and other` $N_2\{pl\}$

Given an anaphor ana and a common noun antecedent candidate a in $\mathcal{A}_{anaid}$, we instantiate (W1) by substituting N_1 with a and N_2 with *ana*. An instantiated pattern for Example (4) is `(home OR homes) and other facilities` (WI_1^c in Table 4).[25] This pattern instantiation is parallel to the WordNet hyp/syn relation look-up for common nouns.

For NE antecedents we instantiate (W1) by substituting N_1 with the NE category of the antecedent, and N_2 with ana. An instantiated pattern for Example (9) is `(person OR persons) and other children` (WI_1^p in Table 4). In this instantiation, N_1 ("person") is not a hyponym of N_2 ("child"), instead N_2 is a hyponym of N_1 (see the discussion on inverted queries in Section 4.4.1.) Therefore, we also instantiate (W1) by substituting N_1 with ana, and N_2 with the NE type of the antecedent (WI_2^p in Table 4). Finally, for NE antecedents, we use an additional pattern:

(W2) N_1 `and other` $N_2\{pl\}$

which we instantiate by substituting N_1 with the original NE antecedent and N_2 with ana (WI_3^p in Table 4). The three instantiations for NEs are parallel to the three hyp/syn relation look-ups in the WordNet experiment in Section 4.4.1. We submit these instantiations as queries to the Google search engine making use of the Google API technology.

BNC For BNC patterns and instantiations, we exploit its part-of-speech tagging. On the one hand, we restrict the instantiation of N_1 and N_2 to nouns to avoid noise, and, on the other hand, we allow occurrence of modification to improve coverage. We therefore extend (W1) and (W2) to the patterns (B1) and (B2).[26] An instantiation for (B1), e.g., also matches "homes and *the* other *four* facilities". Otherwise the instantiations are produced

23 `http://www.google.com`, estimate from November 2004.
24 In all patterns and instantiations in this paper, "OR" is the boolean operator, "N_1" and "N_2" are variables, "and" and "other" are constants.
25 All common noun instantiations are marked by a superscript "c" and all proper noun instantiations by a superscript "p".
26 The star operator indicates zero or more occurrences of a variable. The variable D can be instantiated by any determiner; the variable A can be instantiated by any adjective or cardinal number.

parallel to the Web (see Table 4). We search the instantiations in the BNC using the IMS Corpus Query Workbench (Christ, 1995).

(B1) `(`$N_1\{sg\}$ `OR` $N_1\{pl\}$`) and D* other A*` $N_2\{pl\}$
(B2) N_1 `and D* other A*` $N_2\{pl\}$

Table 4
Patterns and Instantiations for *other*-anaphora.

COMMON NOUN PATTERNS	COMMON NOUN INSTANTIATIONS
W1: `(`$N_1\{sg\}$ `OR` $N_1\{pl\}$`) and other` $N_2\{pl\}$	WI_1^c: (home OR homes) and other facilities
B1: `(...) and D* other A*` $N_2\{pl\}$	BI_1^c: (home OR homes) and D* other A* facilities
PROPER NOUN PATTERNS	**PROPER NOUN INSTANTIATIONS**
W1: `(`$N_1\{sg\}$ `OR` $N_1\{pl\}$`) and other` $N_2\{pl\}$	WI_1^p: (person OR persons) and other children WI_2^p: (child OR children) and other persons
W2: N_1 `and other` $N_2\{pl\}$	WI_3^p: Will Quinlan and other children
B1: `(...) and D* other A*` $N_2\{pl\}$	BI_1^p: (person OR persons) and D* other A* children BI_2^p: (child OR children) and D* other A* persons
B2: N_1 `and D* other A*` $N_2\{pl\}$	BI_3^p: Will Quinlan and D* other A* children

For both algoWeb and algoBNC, each antecedent candidate a in $\mathcal{A}_{anaid}$ is assigned a score. The procedure, using the notation for the Web, is as follows. We obtain the raw frequencies of all instantiations a occurs in (WI_1^c for common nouns, or WI_1^p, WI_2^p, WI_3^p for proper names) from the Web, yielding $freq(WI_1^c)$, or $freq(WI_1^p)$, $freq(WI_2^p)$ and $freq(WI_3^p)$. The maximum WM_a over these frequencies is the score associated with each antecedent (given an anaphor ana), which we will also simply refer to as the antecedent's Web score. For the BNC, we call the corresponding maximum score BM_a, and refer to it as the antecedent's BNC score. This simple maximum score is biased towards antecedent candidates whose head nouns occur more frequently overall. In a previous experiment we used mutual information to normalise Web scores (Markert, Nissim, and Modjeska, 2003). However, the results achieved with normalised and non-normalised scores showed no significant difference. Other normalisation methods might yield significant improvements over simple maximum scoring and can be explored in future work.

4.5.2 Descriptive Statistics Table 5 gives descriptive statistics for the Web and BNC score distributions for correct/lenient antecedents and distractors, including the minimum and maximum score, mean score and standard deviation, median, and the number of zero scores, scores of 1 and scores greater than 1.
Web scores resulting from simple pattern-based search produce on average significantly higher scores for correct/lenient antecedents (mean: 2416.68/807.63; median: 68/68.5) than for distractors (mean: 290.97; median: 1). Moreover, the method produces significantly fewer zero scores for correct/lenient antecedents (19.6%/22.5%) than for distractors (42.3%).[27] Therefore the pattern-based Web method is a good candidate for

27 Difference in means was calculated via a t-test; for medians we used χ^2, and for zero counts a t-test for proportions. The significance level used was 5%.

Table 5
Descriptive Statistics for Web scores and BNC scores for *other*-anaphora

ALL POSSIBLE ANTECEDENTS (TOTAL: 3875)							
	Min-Max	Mean	SD	med	0 scores	1 scores	scores > 1
BNC	0-22	0.07	0.60	0	3714 (95.8%)	109 (2.8%)	52 (1.4%)
Web	0-283 000	542.15	8352.46	2	1513 (39.0%)	270 (7.0%)	2092 (54.0%)
CORRECT ANTECEDENTS (TOTAL: 419)							
	Min-Max	Mean	SD	med	0 scores	1 scores	scores > 1
BNC	0-22	0.32	1.62	0	360 (85.9%)	39 (9.3%)	20 (4.8%)
Web	0-283 000	2416.68	15947.93	68	82 (19.6%)	11 (2.6%)	326 (77.8%)
LENIENT ANTECEDENTS (TOTAL: 160)							
	Min-Max	Mean	SD	med	0 scores	1 scores	scores > 1
BNC	0-4	0.21	0.62	0	139 (86.9%)	13 (8.1%)	8 (5.0%)
Web	0-8840	807.63	1718.13	68.5	36 (22.5%)	3 (1.9%)	121 (75.6%)
DISTRACTORS (TOTAL: 3296)							
	Min-Max	Mean	SD	med	0 scores	1 scores	scores > 1
BNC	0-6	0.03	0.25	0	3215 (97.5%)	57 (1.7%)	24 (0.8%)
Web	0-283 000	290.97	7010.07	1	1395 (42.3%)	256 (7.8%)	1645 (49.9%)

distinguishing correct/lenient antecedents and distractors in anaphora resolution. In addition, the median for correct/lenient antecedents is relatively high (68/68.5), which ensures a relatively large amount of data upon which to base decisions. Only 19.6% of correct antecedents have zero-scores, which indicates that the method might have high coverage (compared to the missing 57% of hyp/syn relations for correct antecedents in WordNet; Section 4.4).

Although the means of the BNC score distributions of correct/lenient antecedents are significantly higher than the one of the distractors, this is due to a few outliers; more interestingly, the median for the BNC score distributions is 0 for all antecedent groups. This will affect precision for a BNC-based algorithm because of the small amount of data decisions are based on. In addition, although the number of zero scores for correct/lenient antecedents (85.9%/86.9%) is significantly lower than for distractors (97.5%), the number of zero scores is well above 80% for all antecedent groups. Thus, the coverage and recall of a BNC-based algorithm will be very low. Although the BNC scores are in general much lower than Web scores and although the Web scores distinguish better between correct/lenient antecedents and distractors, we observed that Web and BNC scores still correlate significantly with correlation coefficients between 0.20 and 0.35, depending on antecedent group.[28]

To summarise, the pattern-based method yields correlated results on different corpora, but it is expected to depend on large corpora to be really successful.

4.5.3 The corpus-based algorithms The prototype Web-based algorithm resolves each anaphor ana to the antecedent candidate in $\mathcal{A}_{anaid}$ with the highest Web score above zero. If several potential antecedents achieve the same Web score, it uses a tiebreaker based on string match and recency. If no antecedent candidate achieves a Web score above 0, string match and recency can be used as a backoff. String comparison for tiebreaker and backoff can again use the original or the replaced antecedents, yielding two versions $\mathsf{algoWeb}_{v1}$ (original antecedents) and $\mathsf{algoWeb}_{v2}$ (replaced antecedents).

28 Correlation significance was measured by both a t-test for the correlation coefficient and also by the non-parametric paired Kendall rank correlation test, both yielding significance at the 1% level.

The exact procedure for the version algoWeb$_{v1}$ for an anaphor ana is as follows:[29]

(i) for each antecedent a in $\mathcal{A}_{anaid}$, compute its Web score WM_a. Compute the maximum WM of all Web scores over all antecedents in $\mathcal{A}_{anaid}$. If WM_a is equal to WM and bigger than zero, push a into a set $\mathcal{A}^{WM}_{anaid}$;
(ii) if $\mathcal{A}^{WM}_{anaid}$ contains exactly one element, select this element and stop;
(iii) otherwise, if $\mathcal{A}^{WM}_{anaid}$ contains more than one element, string-compare each antecedent in $\mathcal{A}^{WM}_{anaid}$ with ana (using original antecedents). If exactly one element of $\mathcal{A}^{WM}_{anaid}$ matches ana, select this one and stop; if several match ana, select the closest to ana within these matching antecedents and stop; if none match, select the closest to ana within $\mathcal{A}^{WM}_{anaid}$ and stop;
(iv) otherwise, if $\mathcal{A}^{WM}_{anaid}$ is empty, make no assigment and stop.

The backoff algorithm algoWeb$^*_{v1}$ uses baselineSTR$^*_{v1}$ as a backoff ([iv']) if no antecedent could be assigned (parallel to the backoff in algoWN$^*_{v1}$):

(iv') otherwise, if $\mathcal{A}^{WM}_{anaid}$ is empty, use baselineSTR$^*_{v1}$ to assign an antecedent to ana and stop;

algoWeb$_{v1}$ and algoWeb$_{v2}$ can overrule string matching for anaphors in StrSet$_{v1}$/StrSet$_{v2}$. This happens when the Web score of an antecedent candidate that does not match the anaphor is higher than the Web scores of matching antecedent candidates. In particular, there is no guarantee that matching antecedent candidates are included in $\mathcal{A}^{WM}_{anaid}$. In that respect, algoWeb$_{v1}$ and algoWeb$_{v2}$ differ from the corresponding WordNet algorithms: matching antecedent candidates are always synonyms of the anaphor (as each noun is a synonym of itself) and therefore always included in $\mathcal{A}^{hyp/syn}_{anaid}$. Therefore the WordNet algorithms can be seen as a direct extension of baselineSTR, i.e. they achieve the same results as the string matching baseline on the sets StrSet$_{v1}$/StrSet$_{v2}$.

Given the high precision of baselineSTR, we might want to exclude the possibility that the Web algorithms overrule string matching. Instead we can use string matching *prior* to Web scoring, only use the Web scores when there are no matching antecedent candidates and use recency as the final backoff. This variation then achieves the same results on the sets StrSet$_{v1}$/StrSet$_{v2}$ as the WordNet algorithms and the string matching baselines. In combination with the possibility of using original or replaced antecedents for string matching this yields 4 algorithm variations overall (see Table 6). The results (see Table 7) do not show any significant differences according to the variation explored.

Table 6
Properties of the variations for the corpus-based algorithms for *other*-anaphora.

	replaced/orig ante	overrule string matching?
v1	orig	yes
v2	replaced	yes
v3	orig	no
v4	replaced	no

The BNC-based algorithms follow the same procedures as the Web-based algorithms, using the BNC scores instead of Web scores. The results (see Table 8) are disappointing

29 The algorithm algoWeb$_{v2}$ follows the same basic procedure apart from the variation regarding original/replaced antecedents in string matching.

Table 7
Web results for *other*-anaphora.

algorithm	coverage	precision	recall	F-measure	$precision^*$
algoWeb$_{v1}$	0.950	0.520	0.495	0.507	0.512
algoWeb$_{v2}$	0.950	0.518	0.493	0.505	0.509
algoWeb$_{v3}$	0.958	0.534	0.512	0.523	0.519
algoWeb$_{v4}$	0.961	0.538	0.517	0.527	0.524

due to data sparseness (see above). No variation yields considerable improvements over baselineSTR$_{v2}$ in the final $precision^*$; in fact, in most cases they just apply a string matching baseline either as a backoff or prior to checking BNC scores, depending on the variation used.

Table 8
BNC results for *other*-anaphora.

algorithm	coverage	precision	recall	F-measure	$precision^*$
algoBNC$_{v1}$	0.210	0.488	0.103	0.170	0.355
algoBNC$_{v2}$	0.210	0.488	0.103	0.170	0.360
algoBNC$_{v3}$	0.417	0.618	0.257	0.363	0.370
algoBNC$_{v4}$	0.419	0.626	0.262	0.369	0.375

4.6 Discussion and Error Analysis

The performances of the best versions of all algorithms for *other*-anaphora are summarised in Table 9.

Table 9
Overview of the results for the best algorithms for *other*-anaphora.

algorithm	coverage	precision	recall	F-measure	$precision^*$
baselineREC	1.000	0.178	0.178	0.178	0.178
baselineSTR$_{v2}$	0.309	0.698	0.216	0.329	0.350
algoBNC$_{v4}$	0.419	0.626	0.262	0.369	0.375
algoWN$_{v2}$	0.652	0.568	0.370	0.448	0.444
algoWeb$_{v4}$	0.961	0.538	0.517	0.527	0.524

Algorithm Comparison Algorithms are compared on their final $precision^*$ using two tests throughout this paper. We used a t-test to measure the difference in the proportion of correctly resolved anaphors between two algorithms. However, there are many examples which are easy (for example, string matching examples) and that therefore most or all algorithms will resolve correctly as well as many that are too hard for all algorithms. Therefore, we also compare two algorithms using McNemar's test, which only

relies on the part of the dataset where the algorithms do not give the same answer.[30] If not otherwise stated, all significance claims hold at the 5% level for both the t-test *and* McNemar's test.

The algorithm baselineSTR significantly outperforms baselineREC in $precision^*$, showing that the "same predicate match" is quite accurate even though not very frequent (coverage is only 30.9%). The WordNet-based and Web-based algorithms achieve a final precision that is significantly better than the baselines' as well as algoBNC's. Most interestingly, the Web-based algorithms significantly outperform the Wordnet-based algorithms, confirming our predictions based on the descriptive statistics. The Web approach, for example, resolves the Examples (3), (4), (6) and (11) in this paper (which WordNet could not resolve) in addition to Examples (8) and (9), which both the Web and WordNet algorithms could resolve.

As expected, the WordNet-based algorithms suffer from the problems discussed in Section 2.2. In particular, Problem 1 proved to be quite severe, as algoWN achieved a coverage of only 65.2%. Missing links in WordNet also affect precision if a good distractor has a link to the anaphor in WordNet whereas the correct antecedent has not (Example (10)). Missing links are both universal relations that should be included in an ontology (such as home:facility) and context-dependent links (e.g. age:(risk) factor, costs:repercussions; see Problem 2). Further mining of WordNet beyond following hyponymy/synonymy links might alleviate Problem 1 but is more costly and might lead to false positives (Problem 3). To a lesser degree, the WordNet algorithms also suffer from sense proliferation (Problem 4), as all senses of both anaphor and antecedent candidates were considered. Therefore, some hyp/syn relations based on a sense not intended in the text were found, leading to wrong antecedent selection and lowering precision. In Example (11), for instance, there is no hyponymy link between the head noun of the correct antecedent ("question") and the head noun of the anaphor ("issue") whereas there is a hyponymy link between "issue" and "person=[Mr. Dallara]" (using the sense of issue as "offspring") as well as a synonymy link between "number" and "issue". While in this case considering the most frequent sense of the anaphor "issue" as indicated in WordNet would help, this would backfire in other cases in our dataset where "issue" is mostly used in the minority sense of "stock, share". Obviously, prior word sense disambiguation would be the most principled but also a more costly solution.

(11) While Mr. Dallara and Japanese officials say *the question of investors access to the U.S. and Japanese markets* may get a disproportionate share of the public's attention, a number of **other important economic issues** [...]

The Web-based method does not suffer as much from these problems. The linguistically motivated patterns we use reduce long-distance dependencies between anaphor and antecedent to local dependencies. By looking up these patterns on the Web we make use of a large amount of data that is very likely to encode strong semantic links via these local dependencies and to do so frequently. This holds both for universal hyponymy relations (addressing Problem 1) and relations that are not necessarily to be included in an ontology (addressing Problem 2). The problem of whether to include subjective and context-dependent relations in an ontology (Problem 2) is circumvented by using Web scores only in comparison to Web scores of other antecedent candidates. In addition, the Web-based algorithm needs no hand-processing or hand-modelling whatsoever, thereby avoiding the manual effort of building ontologies. Moreover, the local dependencies we use reduce the need for prior word sense disambiguation (Problem 4),

30 We thank one anonymous reviewer for suggesting to use McNemar's test for this paper.

as the anaphor and the antecedent constrain each other's sense within the context of the pattern. Furthermore, the Web scores are based on frequency, which biases the Web-based algorithms towards frequent senses as well as sense pairs that occur together frequently. Thus, the Web algorithm has no problem to resolve "issue" to "question" in Example (11) due to the high frequency of the query *question OR questions and other issues*. Problem 3 is still not adressed, however, as any corpus can encode the same semantic relations via different patterns. Combining patterns might therefore yield similar problems as combining information sources in an ontology.

Our pattern-based method, though, seems to work on very large corpora only. Differently from the Web-based algorithms, the BNC-based ones make use of POS tagging and observe sentence boundaries, thus reducing the noise intrinsic to an unprocessed corpus like the Web. Moreover, the instantiations used in algoBNC allow for modification to occur (see Table 4), thus increasing chances of a match. Nevertheless, the BNC-based algorithms performed much worse than the Web: only 4.2% of all pattern instantiations were found in the BNC, yielding very low coverage and recall (see Table 5).

Error Analysis Although the Web algorithms perform best, algoWEB$_{v4}$ still incurs 194 errors (47.6% of 408). Because in several cases there is more than one reason for a wrong assignment, we use the decision tree in Figure 1 for error classification. This way, for example, we can exclude from further analysis those cases that none of the algorithms could resolve because of their intrinsic design.

if all correct/lenient antecedents have been deleted during preprocessing (pronouns) or if the anaphor has split antecedents **then** classify as *design error*;

else if the correct/lenient antecedent is a named entity and has been wrongly resolved or left unresolved by the NER module and this is the cause of the wrong antecedent selection, **then** classify as *NE error*;

else if there is a successful string match to one or more distractors but not to a correct/lenient antecedent, **then** classify as *string matching error*;

else if the correct/lenient antecedent achieves a zero score, **then** classify as *zero score error*;

else if a distractor is selected after winning a tiebreaker against the correct/lenient antecedent, **then** classify as *tiebreaker error*;

else classify as *other error*.

Figure 1
Decision tree for error classification.

As can be seen in Table 10, a quite large number of errors result from deleting pronouns as well as not dealing with split antecedents (44 cases or 22.7% of all mistakes).[31] Out of these 44, 30 involve split antecedents. In 19 of these 30 cases, one of the several correct antecedents has indeed been chosen by our algorithm, *but* all the correct antecedents need to be found to allow for the resolution to be counted as correct.

Given the high number of NE antecedents in our corpus (43.8% of correct, 25% of all antecedents, see Table 1), NE resolution is crucial. In 11.3% of the cases, the algorithm selects a distractor instead of the correct antecedent because the NER module either

31 Percentages of errors are rounded to the first decimal; rounding errors account for the coverage of 99.9% of errors instead of 100%.

Table 10
Occurrences of error types for the best *other*-anaphora algorithm algoWeb$_{v4}$.

ERROR TYPE	# OF CASES	% OF CASES
design	44	22.7
NE	22	11.3
string matching	19	9.8
zero score	48	24.7
tiebreaker	13	6.7
other	48	24.7
total	194	99.9

leaves the correct antecedent unresolved (which could then lead to very few or zero hits in Google) or resolves the named entity to the wrong NE category. String matching is a minor cause of errors (under 10%). This is also due to the fact that there is a possible string match only in just about 30% of the cases (see Table 2).

Many mistakes, instead, are due to the fact that *other-anaphora* can express heavily context-dependent and very unconventional relations, such as the description of "dolls" as "winners" in Example (12).

(12) Coleco bounced back with the introduction of the *Cabbage Patch dolls* [...]. But as the craze died, Coleco failed to come up with **another winner** [...].

In such cases, the relation between the anaphor and antecedent head nouns is not frequent enough to be found in a corpus even as large as the Web.[32] This is mirrored in the high percentage of *zero score* errors (24.7% of all mistakes). Although the Web algorithm suffers from a knowledge gap to a smaller degree than WordNet, there is still a substantial number of cases where we cannot find the right lexical relation.

Errors of type *other* are normally due to good distractors that achieve higher Web scores than the correct antecedent. A common reason is that the wished-for relation is *attested but rare* and therefore other candidates yield higher scores. This is similar to *zero score errors*. Furthermore, the *elimination of modification*, although useful to reduce data sparseness, can sometimes lead to the elimination of information that could help disambiguate among several candidate antecedents. Lastly, lexical information, albeit crucial and probably more important than syntactic information (Modjeska, 2002), is not sufficient for the resolution of *other*-anaphora. The integration of other features, such as grammatical function, NP form and discourse structure, could probably help when very good distractors cannot be ruled out by purely lexical methods (Example (10)). The integration of the Web feature in a machine learning algorithm using several other features has yielded good results (Modjeska, Markert, and Nissim, 2003).

5 Case Study II: definite NP coreference

The Web-based method we have described outperforms WordNet as a knowledge source for antecedent selection in *other*-anaphora resolution. However, it is not clear in how far the method and the achieved comparative results generalise to other kinds of full NP anaphora. In particular, we are interested in the following questions:

32 Using different or simply more patterns might yield some hits for anaphor/antecedent pairs that return a zero score when instantiated in the pattern we use in this paper.

- Is the knowledge gap encountered in WordNet for *other*-anaphora equally severe for other kinds of full NP anaphora? A partial (mostly affirmative) answer to this is given by previous researchers, who put the knowledge gap for coreference at about 30-50% and for bridging at 38-80%, depending on language, domain and corpus (see Section 2).

- Do the Web-based method and the specific search patterns we use generalise to other kinds of anaphora?

- Do different anaphoric phenomena require different lexical knowledge sources?

As a contribution, we investigate the performance of the knowledge sources discussed for *other*-anaphora in the resolution of *coreferential NPs with full lexical heads*, concentrating on definite NPs (see Example (1)). The automatic resolution of such anaphors has received quite significant interest in the past years, but results are much less satisfactory than those obtained for the resolution of pronouns (see Section 2).

The relation between the head nouns of coreferential definite NPs and their antecedents is again, in general, one of hyponymy or synonymy, making an extension of our approach feasible. However, *other*-anaphors are especially apt at conveying context-specific or subjective information by forcing the reader via the *other*-expression to accommodate specific viewpoints. This might not hold for definite NPs.[33]

5.1 Corpus Collection

We extracted definite NP anaphors and their candidate antecedents from the MUC-6 coreference corpus, including both the original training and test material, for a total of 60 documents. The documents were automatically preprocessed in the following way: all meta-information about each document indicated in XML (such as WSJ category and date) was discarded; the headline was included and counted as one sentence. Whenever headlines contained three dashes ("—"), everything after the dashes was discarded.

We then converted the MUC coreference chains into an anaphor-antecedent annotation concentrating on anaphoric definite NPs. All definite NPs which are in, but not at the beginning of, a coreference chain are potential anaphors. We excluded definite NPs with proper noun heads (such as "the United States") from this set since these do not depend on an antecedent for interpretation and are therefore not truly anaphoric.[34] We also exclude appositives, which provide coreference structurally and are therefore not anaphoric. Otherwise, we strictly followed the MUC annotation for coreference in our extraction, although it is not entirely consistent and not necessarily comprehensive (van Deemter and Kibble, 2000). This extraction method yielded a set of 565 anaphoric definite NPs.

For each extracted anaphor in a coreference chain C we regard the NP in C that is closest to the anaphor as the correct antecedent, whereas all other previous mentions in C are regarded as lenient. NPs that occur before the anaphor but are not marked as being in the same coreference chain are *distractors*. Since anaphors with split antecedents are not annotated in MUC, anaphors cannot have more than one correct antecedent. In Example (13), the NPs with the head nouns "Pact" "contract" and "settlement" are marked as coreferent in MUC: in our annotation, "the settlement" is an anaphor with

33 We thank one anonymous reviewer for pointing out that this role for coreference is more likely to be provided by demonstratives than definite NPs.

34 Proper noun heads are approximated by capitalisation in the exclusion procedure.

a correct antecedent headed by "contract" and a lenient antecedent "Pact". Other NPs prior to the anaphor (e.g. "Canada" or "the IWA-Canada union") are distractors.[35]

(13) Forest Products Firms Tentatively Agree On Pact in Canada. A group of large British Columbia forest products companies has reached *a tentative, three-year labor contract with about 18,000 members of the IWA-Canada union*, [...] **The settlement** involves [...]

With respect to *other*-anaphora, we expanded our window size from 2 to 5 sentences (the current and the 4 previous sentences) and excluded all anaphors with no correct or lenient antecedent within this window size, thus yielding a final set of 477 anaphors (84.4% of 565). This larger window size is motivated by the fact that a window size of 2 would only cover 62.3% of all anaphors (352 out 565).

5.2 Antecedent Extraction, Preprocessing and Baselines

All NPs prior to the anaphor within the 5 sentence window were extracted as antecedent candidates.[36] We further processed anaphors and antecedents as in Case Study I (see Section 4.2): modification was stripped and all NPs were lemmatised. In this experiment, named entities were resolved using Curran and Clark's (2003) NE tagger rather than GATE.[37] The identified named entities were further subclassified into finer-grained entities as described for Case Study I.

The final number of extracted antecedents for the whole dataset of 477 anaphors is 14233, with an average of 29.84 antecedent candidates per anaphor. This figure is much higher than the average number of antecedent candidates for *other*-anaphors (10.5) because of the larger window size used. The dataset includes 473 correct antecedents, 803 lenient antecedents and 12957 distractors. Table 11 shows the distribution of NP types for correct and lenient antecedents and for distractors.

Table 11
Distribution of antecedent NP types for definite NPs anaphora

	CORRECT	LENIENT	DISTRACTORS	ALL
pronouns	70	145	1078	1293
named entities	123	316	3108	3547
common nouns	280	342	8771	9133
total	473	803	12957	14233

There are fewer correct antecedents (473) than anaphors (477) because the MUC annotation also includes anaphors whose antecedent is not an NP but, for example, a nominal modifier in a compound. Thus, in Example (14) "the bankruptcy code" is annotated in MUC as coreferential to "bankruptcy-law", a modifier in " bankruptcy-law protection".

(14) All legal proceedings against Eastern, a unit of Texas Air Corp., were put on hold when Eastern filed for *bankruptcy-law* protection March 9. [...] If it doesn't go quickly enough, the judge said he may invoke a provision of **the bankruptcy code** [...]

35 All examples in the coreference study are from the MUC-6 corpus.
36 This extraction proceeded manually to put this study on an equal footing with Case Study I. It presupposes perfect NP chunking. A further discussion of this issue can be found in Section 6.
37 Curran and Clark's (2003) tagger was not available to us during the first case study. Both NE taggers are state-of-the art taggers trained on newspaper text.

In our scheme we extract "the bankruptcy code" as anaphoric but our method of extracting candidate antecedents does not include "bankruptcy-law". Therefore, there are 4 anaphors in our dataset with no correct/lenient antecedent extracted. These cannot be resolved by any of the suggested approaches.

We use the same evaluation measures as for *other*-anaphora as well as the same significance tests for $precision^*$. We also use the same baseline variations baselineREC, baselineSTR$_{v1}$ and baselineSTR$_{v2}$ (see Table 12 and cf. Table 2). The recency baseline performs worse than for *other*-anaphora. String matching improves dramatically on simple recency. It also seems be more relevant than for our *other*-anaphora dataset, achieving higher coverage, precision and recall. This confirms the high value of string matching that has been assigned to coreference resolution by previous researchers (Soon, Ng, and Lim, 2001; Strube, Rapp, and Mueller, 2002, among others).

As the MUC dataset does not include split antecedents, an anaphor ana usually agrees in number with its antecedent. Therefore, we also explored variations of all algorithms that as a first step delete from $\mathcal{A}_{anaid}$ all candidate antecedents that do not agree in number with ana.[38] The algorithms then proceed as usual. Algorithms that use number checking are marked with an additional n in the subscript. Using number checking leads to small but consistent gains for all baselines.

Table 12
Overview of the results for all baselines for coreference

algorithm	coverage	precision	recall	F-measure	$precision^*$
baselineREC	1.000	0.031	0.031	0.031	0.031
baselineSTR$_{v1}$	0.637	0.803	0.511	0.625	0.532
baselineSTR$_{v2}$	0.717	0.775	0.555	0.647	0.570
with number checking					
baselineREC$_n$	1.000	0.086	0.086	0.086	0.086
baselineSTR$_{v1n}$	0.614	0.833	0.511	0.634	0.549
baselineSTR$_{v2n}$	0.694	0.809	0.562	0.664	0.591

As in Case Study I, we deleted pronouns for the WordNet- and corpus-based methods, thereby removing 70 of 473 (14.8%) of correct antecedents (see Table 11). After pronoun deletion, the total number of antecedents in our dataset is 12940 for 477 anaphors, of which 403 are correct antecedents, 658 are lenient antecedents, and 11879 are distractors.

5.3 WordNet for antecedent selection in definite NP coreference

We hypothesize that again most antecedents are hyponyms or synonyms of their anaphors in definite NP coreference (see Examples (1) and (13)). Therefore we use the same lookup for *hyp/syn relations* that was used for *other*-anaphora (see Section 4.4), including the specifications for common noun and proper name lookups. Parallel to Table 3, Table 13 summarises how many correct and lenient antecedents and distractors stand in a hyp/syn

38 The number feature can have the values *singular, plural* or *unknown*. All NE antecedent candidates received the value *singular* as this was by far the most common occurrence in the dataset. Information about the grammatical number of anaphors and common noun antecedent candidates was calculated and retained as additional information during the lemmatisation process. If lemmatisation to both a plural and a singular noun (as determined by WordNet and CELEX) was possible (for example, the word "talks" could be lemmatised to "talk" or "talks"), the value *unknown* was used. An anaphor and an antecedent candidate were said to agree in number if they have the same value or if at least one of the two values is *unknown*.

relation to their anaphor in WordNet.

Table 13
Descriptive statistics for WordNet hyp/syn relations on the coreference dataset.

	hyp/syn relation to ana	no hyp/syn relation	total
correct antecedents	290 (71.96%)	113 (28.04%)	403 (100%)
lenient antecedents	446 (67.78%)	212 (32.22%)	658 (100%)
distractors	1046 (8.80%)	10833 (91.20%)	11879 (100%)
all antecedents	1782 (13.77%)	11158 (86.23%)	12940 (100%)

As already observed for *other*-anaphora, correct and lenient antecedents stand in a hyp/syn relation to their anaphor significantly more often than distractors do (t-test, $p < 0.001$). Hyp/syn relations in WordNet might be better at capturing the relation between antecedent and anaphors for definite NP coreference than for *other*-anaphora:[39] A higher percentage of correct and lenient antecedents of definite NP coreference (71.96%/67.78%) stand in a hyp/syn relation to their anaphors as is the case for *other*-anaphora (43.0%/42.5%). At the same time, though, there is no difference in the percentage of distractors that stand in a hyp/syn relation to their anaphors (9% for *other*-anaphora; 8.80% for definite NP coreference). For our WordNet algorithms, this is likely to translate directly into higher coverage and recall, and potentially into higher precision than in Case Study I. Still, about 30% of correct antecedents are not in a hyp/syn relation to their anaphor in the current case study, confirming results by Harabagiu, Bunescu, and Maiorano (2001), who also look at MUC-style corpora.[40] This gap, though, gets alleviated by a quite high number of lenient antecedents, whose resolution can make up for a missing link between anaphor and correct antecedent.[41]

The WordNet based algorithms are defined exactly as in Section 4.4, with the additional two algorithms that include number checking. Results are summarised in Table 14.

Table 14
Overview of the results for all WordNet algorithms for coreference.

algorithm	coverage	precision	recall	F-measure	$precision^*$
algoWN$_{v1}$	0.874	0.715	0.625	0.666	0.631
algoWN$_{v2}$	0.874	0.724	0.633	0.676	0.639
With Number Checking					
algoWN$_{v1n}$	0.866	0.734	0.635	0.681	0.648
algoWN$_{v2n}$	0.866	0.751	0.649	0.697	0.662

All variations of the WordNet algorithms perform significantly better than the corresponding versions of the string matching baseline (i.e., algoWN$_{v1}$ is better than baselineSTR$_{v1}$, [...], algoWN$_{v2n}$ is better than baselineSTR$_{v2n}$) showing that they add additional lexical knowledge to string matching. As expected from the descriptive statistics discussed

39 Some of this difference might be due to the corpus used instead of the phenomenon as such.
40 Harabagiu, Bunescu, and Maiorano (2001) include all common noun coreference links in their countings, whereas we concentrate on definite NPs only, so that the results are not exactly the same.
41 The possibility of resolving to lenient antecedents follows a similar approach as (Ng and Cardie, 2002b) who suggest a "best-first" coreference resolution approach instead of a " most recent first" approach.

above, the results are better than those obtained by the WordNet algorithms for *other*-anaphora, even if we disregard the additional morphosyntactic number constraint.

5.4 The corpus-based approach for definite NP coreference

Following the assumption that most antecedents are hyponyms or synonyms of their anaphors in definite NP coreference, we use the same list-context pattern and instantiations that were used for *other*-anaphora, allowing us to evaluate whether they are transferrable. The corpora we use are again the Web and the BNC.

Similar to *other*-anaphora, the Web scores do well in distinguishing between correct/lenient antecedents and distractors, with significantly higher means/medians for correct/lenient antecedents (median 472/617 vs. 2 for distractors) as well as significantly fewer zero scores (8% for correct/lenient vs. 41% for distractors). This indicates transferrability of the web-based approach to coreference. Compared to *other*-anaphora the number of zero-scores is lower for correct/lenient antecedent types, so that we expect better overall results, similar to our expectations for the WordNet algorithm.

The BNC scores can also distinguish between correct/lenient antecedents and distractors since the number of zero scores for correct/lenient antecedents (68.98%/58.05%) is significantly lower than for distractors (96.97%). Although more than 50% of correct/lenient antecedents receive a zero-score, there are fewer zero-scores than for *other-anaphora* (where more than 80% of correct/lenient antecedents received zero-scores). However, BNC scores are again in general much lower than Web scores, as measured by means, medians and zero-scores. Nevertheless, Web scores and BNC scores correlate significantly, reaching higher correlation coefficients (0.53 to 0.65 depending on antecedent group) than they did in the case study for *other*-anaphora.

The corpus-based algorithms for coreference resolution are parallel to those described for *other*-anaphora and are marked by the same subscripts. The variations that include number checking are again marked by a subscript n. Table 15 and Table 16 report the results for all the Web and BNC algorithms, respectively.

Table 15
Overview of the results for all Web algorithms for coreference.

algorithm	coverage	precision	recall	F-measure	$precision^*$
algoWeb$_{v1}$	0.994	0.561	0.558	0.559	0.562
algoWeb$_{v2}$	0.994	0.553	0.549	0.550	0.554
algoWeb$_{v3}$	0.998	0.674	0.673	0.673	0.673
algoWeb$_{v4}$	0.998	0.679	0.677	0.678	0.677
With number checking					
algoWeb$_{v1n}$	0.992	0.613	0.608	0.610	0.612
algoWeb$_{v2n}$	0.992	0.607	0.602	0.604	0.606
algoWeb$_{v3n}$	0.996	0.705	0.702	0.703	0.703
algoWeb$_{v4n}$	0.996	0.716	0.713	0.714	0.713

5.5 Discussion and Error Analysis

Algorithm comparison Using the original or the replaced antecedent for string matching (versions $v1$ vs. $v2$, $v1n$ vs. $v2n$, $v3$ vs. $v4$, and $v3n$ vs. $v4n$) never results in interesting differences in any of the approaches discussed. Also, number matching provides consistent improvements. Therefore, we will from now on disregard the variations using original antecedents only ($v1$, $v1n$, $v3$ and $v3n$) as well as algorithms that do not

Table 16
Overview of the results for all BNC algorithms for coreference.

algorithm	coverage	precision	recall	F-measure	$precision^*$
algoBNC$_{v1}$	0.438	0.559	0.245	0.341	0.524
algoBNC$_{v2}$	0.438	0.559	0.245	0.341	0.526
algoBNC$_{v3}$	0.769	0.749	0.576	0.651	0.589
algoBNC$_{v4}$	0.777	0.757	0.589	0.663	0.599
With number checking					
algoBNC$_{v1n}$	0.411	0.612	0.251	0.356	0.562
algoBNC$_{v2n}$	0.411	0.622	0.256	0.369	0.570
algoBNC$_{v3n}$	0.753	0.769	0.579	0.661	0.610
algoBNC$_{v4n}$	0.761	0.785	0.597	0.678	0.627

use number matching ($v2$, $v4$) in our discussion. We will also concentrate on the final $precision^*$ of the full coverage algorithms. The set of anaphors that are covered by the best string matching baseline, *prior to recency backoff*, will again be denoted by StrSet$_{v2n}$. Again, both a t-test and McNemar;s test were used, when statements about significance are made.

The results for the string matching baselines and for the lexical methods are higher for definite coreferential NPs than for *other*-anaphora. This is largely due to the higher number of string matching antecedent/anaphor pairs in coreference, the higher precision of string matching and, to a lesser degree, to the lower number of unusual redescriptions.

Similar to the results for *other*-anaphora, the WordNet-based algorithms beat the corresponding baselines. The first striking result is that the Web algorithm variation algoWeb$_{v2n}$ that relies on the highest web scores only and is therefore allowed to overrule string matching, does not outperform the corresponding string matching baseline baselineSTR$_{v2n}$ and performs significantly worse than the corresponding WordNet algorithm algoWN$_{v2n}$. This contrasts with the results for *other*-anaphora. Examining the results in detail, it emerged that for a considerable number of anaphors in StrSet$_{v2n}$ the highest Web score was indeed achieved by a distractor with a high-frequency head noun when the correct or lenient antecedent could be instead found by a simple string match to the anaphor. This problem is much more severe than for *other*-anaphora because of a) the larger window size that includes more distractors and b) the higher a priori precision of the string matching baseline, which means that overruling string matching leads to wrong results more frequently. Typical examples involve named entity recognition and inverted queries. Thus, in Example (15), the anaphor "the union" is coreferent with the first occurrence of "the union", a case easily resolved by string matching. However, the distractor "organization [=Chrysler Canada]" achieves a higher web score, due to the score of the inverted query "union OR unions and other organizations".[42]

(15) [. . .] *The union* struck Chrysler Canada Tuesday after rejecting a company offer on pension adjustments. **The union** said the size of the adjustments was inadequate.

42 Remember that this problem does not affect the WordNet-based algorithm, which always achieves the same results as the string matching baseline on StrSet$_{v2n}$. Both the correct antecedent and the "organization [=Chrysler Canada]" distractor stand in a hyp/syn relation to the anaphor, and then string matching is used as a tiebreaker.

Several potential solutions exist to this problem, such as normalization of Web scores or penalising inverted queries. The solution we have adopted in algoWeb$_{v4n}$ is to use web scores only *after* string matching, thereby making the Web-based approach better comparable to the WordNet approach. Therefore, baselineSTR$_{v2n}$, algoWeb$_{v4n}$ and algoWN$_{v2n}$ (as well as algoBNC$_{v4n}$) all coincide in their decisions for anaphors in StrSet$_{v2n}$ and only differ in the decisions taken for anaphors that do not have a matching antecedent candidate. Indeed, algoWeb$_{v4n}$ performs significantly better than the baselines at the 1% level and results rise from a $precision^*$ of 60.6% for algoWeb$_{v2n}$ to 71.3% for algoWeb$_{v4n}$. It also significantly outperforms the best BNC results, thus showing that overcoming data sparseness is more important than working with a controlled, tagged and representative corpus. It also shows better performance than WordNet in the final algorithm variation (71.3% vs. 66.2%).[43] Using a t-test, this last difference is, however, not significant. McNemar's test, concentrating on the part of the data where the methods differ, shows instead significance at the 1% level.

Indeed, one of the problems in comparing algorithm results for coreference is that such a large number of anaphors are covered by simple string matching, leaving only a small dataset on which the lexical methods can differ. Thus, StrSet$_{v2n}$ contains 331 of 477 cases (268 of which are assigned correctly by baselineStr$_{v2n}$) so that improvements by the other methods are confined to the set of the remaining 146 anaphors. Of these 146, baselineStr$^*_{v2n}$ assigns the correct antecedent to 13 (8.9%) anaphors by using a recency backoff, the best WordNet method to 55 (37.67%), and the best Web method to 72 anaphors (49.31%). Therefore the Web-based method is a better complement to string matching than WordNet, which is reflected in the results of McNemar's test. Anaphor-antecedent relations that were not covered in WordNet but that did not prove a problem for the Web algorithm were again both general hyponymy relations such as retailer:organization, bill:legislation and month:time as well as more subjective relations like (wage) cuts:concessions and legislation:attack.

Error Analysis The best performing Web-based algorithm, algoWeb$_{v4n}$, still selects the wrong antecedent for a given anaphor in 137 of 477 cases (28.7%). Again, we use the decision tree in Figure 1 to classify errors. Design errors now do not include split antecedents but do include errors that occur because the condition of number agreement was violated, pronoun deletion errors, and the 4 cases in which the antecedent was a non-NP antecedent and therefore not extracted in the first place (see Section 5.1 and Example (14)). Table 17 reports the frequency of each error type.

Table 17
Occurrences of error types for the best coreference algorithm algoWeb$_{v4n}$.

ERROR TYPE	# OF CASES	% OF CASES
design	12	8.7
NE	7	5.1
string matching	33	24.1
zero scores	11	8.0
tiebreaker	34	24.8
other	40	29.2
total	137	99.9

43 In general, the WordNet methods achieve higher precision, with the Web-method achieving higher recall.

Differently from *other*-anaphora, the design and NE errors together only account for under 15% of the mistakes. Also rare are zero score errors (only 8%). When compared to the number of *zero score* errors in other anaphora (24.7%), this low figure suggests that *other*-anaphora is more prone to exploit rare, unusual and context-dependent redescriptions than full NP coreference. Nevertheless, it is yet possible to find non-standard redescriptions in coreference as well which yield zero scores, such as the use of "transaction" to refer to "move" in Example (16).

(16) Conseco Inc., in *a move* to generate about $200 million in tax deductions, said it induced five of its top executives to exercise stock options to purchase about 3.6 million common shares of the financial-services concern. As a result of **the transaction**, [...]

Much more substantial is the weight of errors due to string matching, tiebreaker decisions, and the presence of good distractors (the main reason for errors of type *other*), which together account for over three quarters of all mistakes.

String matching is quite successful for coreference (baselineSTR$_{v2n}$ covers nearly 70% of the cases with a precision of 80.9%). However, because algoWeb$_{v4n}$ never overrules string matching, the errors of baselineSTR$_{v2n}$ are preserved here, and account for 24.1% of all mistakes.[44] Tiebreaker errors are quite frequent too (24.8%), as our far-from-sophisticated tiebreaker was needed in nearly half of the cases (224 times; 47.0%).

The remaining errors (29.2%) are due to the presence of good distractors that score higher than the correct/lenient antecedent. In Example (17), for instance, a distractor with a higher web score ("comment") prevents the algorithm from selecting the correct antecedent ("investigation") for the anaphor "the inquiry".

(17) Mr. Adams couldn't be reached for comment. Though *the investigation* has barely begun, persons close to the board said Messrs. Lavin and Young will get a "hard look" as to whether they were involved, and are both considered a "natural focus" of **the inquiry**.

Example (18) shows how stripping modification might have eliminated crucial information to identify the correct antecedent: only the head "process" was kept of the anaphor "arbitration process", so that the surface link between anaphor and antecedent ("arbitration") was lost and the distractor "securities industry", reduced to "industry", was instead selected.

(18) The securities industry has favored *arbitration* because it keeps brokers and dealers out of court. But consumer advocates say customers sometimes unwittingly sign away their right to sue. "We don't necessarily have a beef with **the arbitration process**," says Martin Meehan, [...]

6 Open Issues

Preprocessing and prior assumptions Our algorithms build on two main pre-processing assumptions. Firstly, we assume perfect base NP chunking and expect results to be lower with automatic chunking. Nevertheless, since automatic chunking will affect all algorithms in the same way, we do expect comparative results to stand. We are however not dependent on full parsing, as no parsing-dependent grammatical features are used by the algorithms.

44 Some of the errors incured by baselineSTR$_{v2n}$ are here classified as design, NE, or tiebreaker errors.

Secondly, the anaphoricity of the definite NPs in Case Study II has de facto been manually determined as we restrict our study to antecedent selection for the NPs that are marked in the MUC corpus as coreferent. One of the reasons why pronoun resolution has been more successful than definite NP resolution is that whereas pronouns are mostly anaphoric, definite NPs do not have to be so (see Section 2). In fact, it has been argued by several researchers that an anaphora resolution algorithm should proceed to antecedent selection *only if* a given definite NP *is* anaphoric (Ng and Cardie, 2002a; Ng, 2004; Uryupina, 2003; Vieira and Poesio, 2000, among others), therefore advocating a two-stage process which we also follow in this paper. Although recent work on automatic anaphoricity determination has shown promising results (Ng, 2004; Uryupina, 2003), our algorithms will perform worse when building on non-manually determined anaphors. Future work will explore the extent of such a decrease in performance.

Directions for improvement All algorithms we have described can be considered a blueprint for more complex versions. Specifically, the WordNet-based algorithms could be improved by exploiting information encoded in WordNet beyond explicitly encoded links (glosses could be mined, too, for example; see also (Harabagiu, Bunescu, and Maiorano, 2001)). The Web-based algorithms could similarly benefit from the exploration of different patterns and their combination as well as using non-pattern based approaches for hyponymy detection (Shinzato and Torisawa, 2004). In addition, we have evaluated the contribution of lexical resources *in isolation* rather than within a more sophisticated system that integrates additional non-lexical features. It is unclear whether integrating such knowledge sources in a full resolution system might even out the differences between the Web-based and the WordNet based algorithms or exacerbate them. Modjeska, Markert, and Nissim (2003) included a feature based on Web scores in a Naive Bayes model for *other*-anaphora resolution that also used grammatical features, and showed that the addition of the Web feature yielded an 11.4 percentage point improvement over using a WordNet-based feature. This gives some indication that additional grammatical features might not be able to compensate fully for the knowledge gap encountered in WordNet.

Extension to yet other anaphora types Using the Web for antecedent selection in anaphora resolution is novel and needs further study for other types of full NP anaphora than the ones studied in this paper. If an anaphora type exploits hyponymy/synonymy relationships between anaphor and antecedent head nouns, it can in principle be treated with exactly the same pattern we used in this paper. This holds, for example, for demonstratives and *such*-anaphors. The latter, in particular, are similar to *other*-anaphora in that they establish a comparison between the entity they invoke and that invoked by antecedent, and are also easily used to accommodate subjective viewpoints. They should therefore benefit especially from not relying wholly on standard taxonomic links.

Different patterns can be developed for anaphora types that build on non-hyponymy relations. For example, bridging exploits meronymy and/or causal relations (among others). Therefore, patterns that express "part-of" links, for example, such as "X of Y" and genitives, would be appropriate. Indeed, these patterns have been recently used in Web search for antecedent selection for bridging anaphora by Poesio et al. (2004). They compare accuracy in antecedent selection for a method that integrates Web hits and focusing techniques with a method that uses WordNet and focusing, achieving comparable results for both methods. This strenghtens our hypothesis that antecedent selection for full NP anaphora without hand-modelled lexical knowledge has become feasible.

7 Conclusions

We have explored two different ways of exploiting lexical knowledge for antecedent selection in *other*-anaphora and definite NP coreference. Specifically, we have compared a hand-crafted and structured source of information such as WordNet and a simple and inexpensive pattern-based method operating on corpora. As corpora we have used the BNC and also suggested the Web as the biggest corpus available.

We confirmed results by other researchers that show that a substantial number of lexical links often exploited in coreference are not included in WordNet. We have also shown the presence of an even more severe knowledge gap for *other*-anaphora (see also Question 1 in the Introduction). Largely due to this knowledge gap, the novel Web-based method that we proposed proved *better* than WordNet at resolving *other*-anaphora. Although the gains for coreference are not as high, the Web-based method improves more substantially on string matching techniques for coreference than WordNet does (see the success rate beyond StrSet_{v2n} for coreference; Section 5.5). In both studies, the Web-based method clearly outperformed the BNC-based one. This shows that, for our tasks, overcoming data sparseness was more important than working with a manually controlled, virtually noise-free but relatively small corpus. So, this adresses Question 2 in the Introduction: corpus-induced knowledge can indeed rival and even outperform the knowledge obtained via lexical hierarchies, as long as the corpus is large enough. The corpus-based methods can therefore be a very useful complement to resolution algorithms for languages for which hand-crafted taxonomies have not yet been created but for which large corpora do exist. In answer to Question 3 in the Introduction, our results suggest that different anaphoric phenomena suffer in varying degree from missing knowledge and that the Web-based method performs best when used to deal with phenomena that standard taxonomy links do not capture that easily or that frequently exploit subjective and context-dependent knowledge.

In addition, the Web-based method that we propose does not suffer from some of the intrinsic limitations of ontologies, specifically the problem of what knowledge should be included (see Section 2.2). It is also inexpensive, does not need any postprocessing of the web pages returned nor any hand-modelling of lexical knowledge.

To summarize, antecedent selection for *other*-anaphora and definite NP coreference without hand-crafted lexical knowledge is feasible. This might also be the case for yet other full NP anaphora types with similar properties — an issue that we will explore in future work.

Acknowledgments

We especially thank Natalia Modjeska for providing us with her annotated corpus of *other*-anaphors as well as with the extracted and partially preprocessed sets of candidate antecedents for Case Study I. She also collaborated on previous related work on *other*-anaphora (Markert, Nissim, and Modjeska, 2003; Modjeska, Markert, and Nissim, 2003), which this paper builds on. We would also like to thank James Curran, Bonnie Webber and four anonymous reviewers for helpful comments, which allowed us to strongly improve this paper. Malvina Nissim was partially supported by Scottish Enterprise Stanford-Link Grants R36766 (Paraphrase Generation) and R36759 (SEER).

References

Ariel, Mira. 1990. *Accessing Noun Phrase Antecedents*. Routledge, London-New York.

Berland, Matthew and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics; Providence, Rhode Island*, pages 57–64.

Bierner, Gann. 2001. Alternative phrases and natural language information retrieval. In *Proc. of the 39th Annual Meeting of the Association for Computational Linguistics; Toulouse, France*, pages 58–65.

Burnard, Lou, 1995. *Users' Reference Guide, British National Corpus*. British National Corpus Consortium, Oxford, England.

Caraballo, Sharon. 1999. Automatic acquisition of a hypernym-labelled noun hierarchy from text. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics; Providence, Rhode Island*, pages 120–126.

Chinchor, Nancy. 1997. MUC-7 Named Entity Task definition. In *Proc. of the 7th Conference on Message Understanding; 1997*, Washington, DC.

Christ, Oliver, 1995. *The XKWIC User Manual*. Institute for Computational Linguistics, University of Stuttgart.

Clark, Herbert H. 1975. Bridging. In *Proc. of the Conference on Theoretical Issues in Natural Language Processing; Cambridge, Mass.*, pages 169–174.

Connolly, Dennis, John D. Burger, and David S. Day. 1997. A machine learning approach to anaphoric reference. In Daniel Jones and Harold Somers, editors, *New Methods in Language Processing*. UCL Press, London, pages 133–144.

Curran, James and Stephen Clark. 2003. Language Independent NER using a Maximum Entropy Tagger. In *Proc. of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167, Edmonton, Canada, 2003.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Fraurud, Kari. 1990. Definiteness and the processing of NPs in natural discourse. *Journal of Semantics*, 7:395–433.

Gardent, Claire, Helene Manuelian, and Eric Kow. 2003. Which bridges for bridging definite descriptions? In *Proc. of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69–76.

Grefenstette, Gregory. 1999. The WWW as a resource for example-based MT tasks. In *Proceedings of ASLIB'99 Translating and the Computer 21*, London.

Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Hahn, Udo, Michael Strube, and Katja Markert. 1996. Bridging textual ellipses. In *Proc. of the 16th International Conference on Computational Linguistics; Copenhagen, Denmark*, pages 496–501.

Halliday, Michael A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Harabagiu, Sanda. 1997. *WordNet-Based Inference of Textual Context, Cohesion and Coherence*. Ph.D. thesis, University of Southern California.

Harabagiu, Sanda, Razvan Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *Proc. of the 2nd Conference of the North American Chapter of the ACL; Pittsburgh, PA*, pages 55–62.

Hawkins, John A. 1978. *Definiteness and Indefiniteness*. Croom Helm, London.

Hearst, Marti. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics; Nantes, France*.

Hirschman, Lynette and Nancy Chinchor. 1997. MUC-7 coreference task definition. In *Proc. of the 7th Conference on Message Understanding; 1997*.

Humphreys, Kevin, Robert Gaizauskas, Saliha Azzam, Chris Huyck, Brian Mitchell, and Hamish Cunningham. 1997. University of Sheffield: description of the LaSie-II system as used for MUC-7. In *Proc.of the 7th Message Understanding Conference (MUC-7)*.

Kameyama, Megumi. 1997. Recognizing referential links: an information extraction perspective. In *Proc. of the ACL-1997 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53, Madrid, Spain.

Keller, Frank and Maria Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proc. of the 16th International Conference on Computational Linguistics; Copenhagen, Denmark*, pages 113–118.

Markert, Katja, Malvina Nissim, and Natalia N. Modjeska. 2003. Using the web for nominal anaphora resolution. In Robert Dale, Kees van Deemter, and Ruslan Mitkov, editors, *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*, pages 39–46.

McCoy, Kathleen and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *ACL-99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 63–71.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography. In D Bourigault,

C Jacquemin, and M L'Homme, editors, *Recent Advances in Computational Terminology*. John Benjamins, Amsterdam, pages 279–301.

Meyer, Josef and Robert Dale. 2002. Mining a corpus to support associative anaphora resolution. In *Proc. of the Fourth Discourse Anaphora and Anaphor Resolution; 2002*.

Mitkov, Ruslan. 1998. Robust pronoun resolution with limited knowledge. In *Proc. of the 17^{th} International Conference on Computational Linguistics and 36^{th} Annual Meeting of the Association for Computational Linguistics; Montréal, Canada*, pages 869–879.

Modjeska, Natalia N. 2002. Lexical and grammatical role constraints in resolving *other*-anaphora. In *Proc. of DAARC 2002*, pages 129–134, Lisbon, Portugal.

Modjeska, Natalia N. 2003. *Resolving* other-*anaphora*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Modjeska, Natalia N., Katja Markert, and Malvina Nissim. 2003. Using the web in machine learning for other-anaphora resolution. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing; Sapporo, Japan*, pages 176–183.

Ng, Vincent. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proc. of the 42^{nd} Annual Meeting of the Association for Computational Linguistics; Barcelona, Spain; 2004*, pages 151–158.

Ng, Vincent and Claire Cardie. 2002a. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proc of the 19^{th} International Conference on Computational Linguistics; Taipei, Taiwan*, pages 730–736.

Ng, Vincent and Claire Cardie. 2002b. Improving machine learning approaches to coreference resolution. In *Proc. of the 40^{th} Annual Meeting of the Association for Computational Linguistics; Philadelphia, Penn.*, pages 104–111.

Poesio, Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Viera. 2002. Acquiring lexical knowledge for anaphora resolution. In *Proc. of the 3^{rd} International Conference on Language Resources and Evaluation; Las Palmas, Canary Islands, 2002*, pages 1220–1224.

Poesio, Massimo, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. 2004. Learning to resolve bridging references. In *Proc. of the 42^{nd} Annual Meeting of the Association for Computational Linguistics; Barcelona, Spain; 2004*, pages 143–150.

Poesio, Massimo, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In R. Mitkov, editor, *Proc. of the ACL Workshop on Operational Factors in Robust Anaphora Resolution*, pages 1–6, Madrid.

Preiss, Judita. 2002. Anaphora resolution with word sense disambiguation. In *Proc. of SENSEVAL-2*, pages 143–146.

Preiss, Judita, Caroline Gasperin, and Ted Briscoe. 2004. Can anaphoric definite descriptions be replaced by pronouns? In *Proc. of the 4^{th} International Conference on Language Resources and Evaluation; Lisbon, Portugal, 2004*, pages 1499–1502.

Shinzato, Keiji and Kentaro Torisawa. 2004. Acquiring hyponymy relations from web documents. In *Proc. of the Conference of the North American Chapter of the ACL; 2004*, pages 73–80.

Soon, Wee Meng, Hwee Tou Ng Ng, and Daniel Chung Yung Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Strube, Michael, Stefan Rapp, and Christoph Mueller. 2002. The influence of minimum edit distance on reference resolution. In *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing; Philadelphia, Penn.*, pages 312–319.

Uryupina, Olga. 2003. High-precision identification of discourse new and unique noun phrases. In *Proc. of the ACL 2003 Student Workshop*, pages 80–86.

van Deemter, Kees and Rodger Kibble. 2000. On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):615–62.

Vieira, Renata and Massimo Poesio. 2000. An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Webber, B., M. Stone, A. Joshi, and A. Knott. 2003. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Yang, Xiaofeng, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proc. of the 41^{st} Annual Meeting of the Association for Computational Linguistics; Sapporo, Japan*, pages 176–183.