Chapter 13
Detecting Non-reference and Non-anaphoricity

Olga Uryupina, Mijail Kabadjov, and Massimo Poesio

Abstract In this Chapter we discuss proposals concerning the detection of non-referentiality and non-anaphoricity, and the integration of such methods in an anaphora resolution system. We first review in brief a number of proposals on the topics, also covering literature on detecting abstract anaphora and discussing available resources; we then discuss in detail the proposals by Bergsma on expletive detection, and by Poesio et al and Kabadjov on discourse-new detection.

13.1 Introduction

As discussed in earlier Chapters, anaphora (or coreference) resolution is the task of identifying the **(discourse) entity** a mention refers to–which in most modern systems, in which entities are represented in terms of **coreference chains** (equivalence sets of mentions referring to the same entity), boils down to partitioning the mentions in a text into these coreference chains. This simple picture is, however, complicated in number of ways. First of all, not all mentions refer. As discussed in Chapter 2, some mentions are semantically vacuous; other mentions, while making a semantic contribution, are not referring. In addition, many referring mentions do not corefer with any other mention, i.e., the coreference chain to which they belong is in fact a singleton set.

Regarding the first point– English syntax requires the subjects of finite clauses to be explicitly realized. As a result, sentences such as *It rains*, which in languages

Olga Uryupina
University of Trento e-mail: uryupina@gmail.com
Mijail Kabadjov
University of Essex, School of Computer Science and Electronic Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom, e-mail: malexa@essex.ac.uk
Massimo Poesio
University of Essex, School of Computer Science and Electronic Engineering, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. e-mail: poesio@essex.ac.uk

such as Italian, Spanish or Japanese would not have a subject, have as a subject an *it* which doesn't realize any argument of the verb; it's only there for syntactic reasons. Such proforms are called **expletives** or **pleonastic**. Being able to recognize such proforms could improve the performance of an anaphora resolution system, at least in principle.

The second complication is that even mentions which do contribute to the logical form of a sentence may not be **referring**. As discussed in Chapter 2, noun phrases can play at least three types of semantic roles in the logical form of a sentence. First of all, noun phrases can refer to an entity: this is the case of noun phrase *John* in *John is a policeman*. Secondly, they can act as predicates: e.g., noun phrase *a policeman* in the example just given is not referring– it is used to express a property of the entity referred to by mention *John*. Thirdly, nominal phrases can act as **quantifiers**. For example, in *No self-respecting Italian likes pineapple on her pizza*, the nominal phrase *No self-respecting Italian* does not refer to any entity: according to modern linguistics, the sentence means that the intersection of the sets of self-respecting Italians and the set of people liking pineapple on their pizza is empty, and the determiner *No* specifies this. Note however that the quantifier *No self-respecting Italian* can nevertheless act as 'antecedent' for the anaphoric pronoun *her* (pronouns whose 'antecedent' is a quantifier, like *her*, are called **bound anaphors**).[1]

The third complexity to take in mind is that many, in fact most, properly referring mentions are **non-anaphoric** or at least do not corefer with entities previously introduced via nominal mentions. The great majority of mentions in most texts are **discourse new**: they introduce new entities in a discourse. And a great many referring mentions are referring to entities not directly introduced via nominal mentions–e.g., in associative anaphora and in reference to abstract objects (see Chapter 2).

To appreciate the extent to which these complexities affect current anaphora resolution systems, consider the following example from the OntoNotes dataset [28, 50]:

(13.1) Why do $[\text{you}]_1$ think only $[\text{lowly soldiers}]_2$ and $[\text{no generals}]_3$ have been convicted? Because $[\text{generals}]_4$ have $[\text{the power}]_5$, $[\text{that}]_6$'s why!

The snippet in (13.1) contains six nominal expressions, that can potentially be partitioned into coreference chains in 203 ways. Some of the candidate partitions, however, can be reliably ruled out at an early stage, based on simple mention-level contextual clues. First, *no generals* (M_3) is a quantifier, thus non referring to any particular set of generals although it can bind pronominal expressions (but such binding relations are typically not annotated in anaphoric / coreference corpora, see Chapter 4). Most importantly, *no generals* should not be deemed as coreferent with *generals* (M_4), even though they share very similar surface forms. Depending on the annotation guidelines, an anaphora resolution system should either classify this nominal

[1] And indeed, the clearest difference between 'anaphora resolution' as specified in theoretical linguistics / psycholinguistics and 'coreference resolution' as specified by the MUC annotation guidelines lies precisely in the treatment of predicative and quantificational NPs. In anaphora resolution predicative NPs have no antecedents, but bound pronouns do; conversely, in coreference resolution the links between bound pronouns and the quantifiers that bound them are not annotated, but those between predicative NPs and their 'antecedents' are. (See Chapter 4.)

as non-referring, or a belonging to a singleton chain. Second, the pronoun *that* (M_6) refers to an **abstract entity** (*generals have the power*) and should therefore not be part of any coreference chain including previously occurred nominal expressions. This is an important clue since most pronouns are anaphoric and therefore a typical coreference resolution system would tend to propose an antecedent for M_6, leading to a spurious link. Finally, *you* (M_1) is a **deictic** expression that refers to the hearer and not to any particle person mentioned in the snippet.

Many anaphoric resolvers do not include any special provision to recognize such cases. But in many cases, non-referring, non-anaphoric and discourse new expressions can be reliably identified using various linguistic cues. It might therefore be beneficial for a coreference resolver to incorporate submodules for these tasks. First, referentiality and anaphoricity detection can be modeled as simple binary classification problems allowing for straightforward implementation. This can help improve the precision level of our coreference resolver: for example, if our referentiality detector classifies with a high confidence *no generals* in (13.1) as non-referential, this mention can be assigned to a singleton chain in the final partition and the coreference system will not introduce such spurious links as {"no generals", "generals"}. Second, when used as pre-filters in a pipelined architecture, such modules might help significantly reduce the pool of candidate mentions and thus considerably increase the performance speed. Thus, by assigning M_1, M_3 and M_6 to singleton chains, we can reduce the total number of all the possible partitions in our example (13.1) from 203 to just 5. This is especially important in the context of state-of-the-art complex models of coreference that are very sensitive to the size of their input problems.

The rest of this Chapter is organized as follows. In Section 13.2 we survey work on the different subtasks of anaphora resolution concerned with detecting non-referentiality and non-anaphoricity, and on integrating such work with anaphora resolution systems. In Section 13.3, we provide a detailed description of a state-of-the-art algorithm for expletive detection [6]. In Section 13.4, we discuss the in-depth investigation of non-anaphoricity in [29].

13.2 Non-Referentiality, Non-Anaphoricity, and Related Tasks

In this Section we look in more detail at the subtasks of anaphora resolution concerned with referentiality and anaphoricity, and survey the solutions proposed, discussing also relevant work in corpus annotation.

13.2.1 Detecting (Non) Referentiality and (Non) Antecedenthood

13.2.1.1 Expletives

As discussed in Chapter 2, linguists have been distinguishing between anaphoric and expletive usages of pronouns for a very long time [54, 60, 61, 26, 27]. One of the most thoroughly covered topics in this context is the expletive usage of English *it*, and extensive analyses of the constructions in which such usages occur have been provided e.g., in [54, 60, 61]. Early computational approaches to pronominal anaphora resolution circumvented the problem by manually excluding expletives from their evaluation datasets (see e.g., [38] for a discussion). Lappin and Leass's algorithm [36], discussed in Chapter 3, used expletive recognition heuristics such as rules for recognising modal adjectives (e.g., 'It is *good* to know') and cognitive verbs (e.g., 'It is *believed* that polar bears...'). In the past decade, several statistical algorithms for automatic identification of expletive *it* have been proposed [22, 9, 6, 5]. We briefly discuss here the proposals by Evans and by Versley *et al.*, and Bergsma's algorithm in more detail in Section 13.3.

Evans

Evans [22] developed a classifier able to classify every occurrence of the pronoun *it* into one of the seven classes **nominal anaphoric**, **clause anaphoric**, **proaction**, **cataphoric**, **discourse topic**, **pleonastic** and **idiomatic/stereotypic**. Each occurrence of *it* is annotated with 35 features, that can be broadly grouped into the following six categories:

1. **Position** of the pronoun in text is recorded in terms of word position in sentence and sentence position in paragraph. (Pronouns at the beginning of paragraphs are unlikely to be anaphoric.)
2. **Surface features** extracted from the text surrounding the pronoun, such as whether the pronoun immediately follows a prepositional word (pleonastic pronouns seldom appear after a prepositional word) or whether the pronoun is followed by complementisers or adjectives (pleonastic pronouns often precede such expressions).
3. **Word lemmas** of preceding and subsequent text within the same sentence (in particular verbs). These are meant to bypass the need for compiling external lists of trigger words like 'weather adjectives' or 'cognitive verbs'.
4. **Part-of-speech** for a window of eight words centred around the pronoun.
5. Certain **grammatical patterns** like 'adjective + noun phrase' (for example, '*It* was obvious [the plan] would work') and 'complementiser + noun phrase' (for example, '*It* was obvious [that the plan] would work').
6. **Proximity** of the pronoun to complementisers, *-ing* verb forms, and prepositions, in terms of number of tokens in between.

Evans then trained an expletive classifier using the implementation of the k nearest neighbour (KNN) algorithm implemented in Tilburg University's Memory Based Learner (TiMBL) [17]. He used for training and testing a data set he created, formed by 77 texts from the SUSANNE and BNC corpora from a diverse set of genres such as politics, science, fiction and journalism. Their corpus consisted of 368,830 words and contained 3171 occurrences of *it* broken down into the seven target classes as follows: 67.93% to nominal anaphora, 0.82% to clause anaphora, 0.06% to proaction, 0.09% to cataphora, 2.08% to discourse topic, 26.77% were pleonastic, and 2.24% to idiomatic/stereotypic constructions.

Evans reports precision and recall figures for each of the seven classes as follows: $F = 77.12\%$ for nominal anaphora, $F = 0\%$ for clause anaphora, $F = 0\%$ for proaction, $F = 0\%$ for cataphora, $F = 2.85\%$ for discourse topic, $F = 71.26\%$ for pleonastic and $F = 1.37\%$ for idiomatic/stereotypic constructions. Evans' classifier has also been integrated with the MARS anaphora resolution system [38] discussed in Chapter 3.

Versley et al

Versley *et al.* [66] argued that tree kernels are particularly suited for identifying expletives, as they are able automatically to identify generalizations about the lexical and structural context around a word such as those encoded in the features used by Evans. In their experiments, they used the BBN Pronoun corpus to train and test an expletive classifier using the Penn Treebank parse trees. On expletive classification Versley *et al.*'s best classifier achieved a performance of $F = 74.36\%$, compared to a result of $F = 49.6\%$ obtained using a baseline classifying all instances of *it* as expletive. They then integrated this classifier in the BART anaphora resolution system, using the output of the expletive detector as an additional feature into their pronoun resolution module. Their system achieved $F = 66.5\%$ for pronouns on the MUC-6 corpus, a small improvement over the performance of the version of BART they used of $F = 66\%$ but the upperbound with perfect expletive classification would be $F = 68.4\%$.

13.2.1.2 Non-referring nominals

The fact that not all nominals are referring has been known in the semantics literature for several decades [31, 58, 70]. One of the best known references in this area is the work by Karttunen [31], who discussed examples such as (13.2):

(13.2) Bill doesn't have [a car].

Sentence (13.2) does not imply the existence of any specific "car". In Karttunen's terms, the NP *a car* does not *establish a discourse referent* and therefore it cannot participate in any coreference chain—none of the alternatives in (13.3) can follow (13.2):

(13.3) A. [It] is black.
B. [The car] is black.
C. [Bill's car] is black.

Karttunen identifies several factors affecting the referential status of NPs, including modality, negation, or non-factive verbs. He argues that an extensive analysis of the phenomenon requires a sophisticated semantic treatment; and indeed numerous theories of semantic representation were proposed to account for such constraints, most notably Discourse Representation Theory or DRT [30], discussed in Chapter 2. Later linguistic studies further investigated what kind of entities are allowed to participate in coreferent chains, and identified some of the relevant factors: internal morphosyntax, interaction between negation, modality, quantification, attitude predicates among others [58, 70].

Quite a lot of anaphora resolution systems based on DRT, thus incorporating its treatment of non-referential NPs, were developed in computational linguistics in the '80s [2], the most recent system of this type being the Boxer semantic interpreter [8]. An algorithm for identifying *nonlicensing* NPs based on Karttunen's theory of referentiality was proposed by Byron and Gegg-Harrison [12]. Their approach relies on a handcrafted heuristic, encoding some of Karttunen's factors and shows mixed results w.r.t. the impact of such a prefiltering on (pronominal) coreference resolution.

13.2.1.3 Pragmatic effects on antecedenthood

Empirical analysis of the results of anaphora resolution has shown that besides the semantic constraints just discussed there are pragmatic restrictions on the likelihood that entities will serve as antecedents for subsequent reference. The term **antecedenthood detection** [63] has been used for the task of identifying mentions that are not likely to be antecedents. Further discussion of the degree of antecedenthood of entities can be found in [55]. A generalization of the antecedenthood and anaphoricity tasks, the **entity lifespan** detection problem, has been proposed recently [56]: the objective is, for a given nominal expression, to predict the size of its corresponding coreference chain. A binary classifier is trained based on contextual features to discriminate between singletons and longer chains. This is an important area of research, that needs further investigation.

13.2.2 Discourse-New Detection and Anaphora Resolution

13.2.2.1 Detecting discourse-new mentions

The task of **discourse new detection** consists of classifying nominal mentions according to their **information status**: i.e., distinguishing between descriptions in-

troducing new entities and those referring to already established ones. Numerous theories concerning the information status of nominals have been proposed in the literature [24, 52, 53, 37, 23]. We point the reader to [65] and [67] for extensive overviews and comparison of some of the main theoretical proposals in this regard. Of these, the best known is the theory developed by Prince [52, 53], who introduces a distinction between between **discourse givenness** and **hearer givenness** that results in the following taxonomy:

- *brand new* NPs introduce entities that are both discourse and hearer new ("a bus"), some of them, *brand new anchored* NPs, contain explicit link to some given discourse entity ("a guy I work with"),
- *unused* NPs introduce discourse new, but hearer old entities ("Noam Chomsky"),
- *evoked* NPs introduce entities already present in the discourse model and thus discourse and hearer old: *textually evoked* NPs refer to entities which have already been mentioned in the previous discourse ("he" in "A guy I worked with says he knows your sister"), whereas *situationally evoked* are known for situational reasons ("you" in "Would you have change of a quarter?"),
- *inferrables* are not discourse or hearer old, however, the speaker assumes the hearer can infer them via logical reasoning from evoked entities or other inferrables ("the driver" in "I got on a bus yesterday and the driver was drunk"), *containing inferrables* make this inference link explicit ("one of these eggs").

Many linguistic theories [24, 52, 37, 23] focus on anaphoric usages of definite descriptions (either evoked or inferrables). As a result, many of the early systems focused exclusively on the task of finding an antecedent for a given expression. Corpus studies have revealed, however, that more than 50% of (definite) NPs in newswire texts are not anaphoric [23, 49]. These findings suggest that developing data-driven approaches for the automatic identification of discourse new NPs may improve the performance of anaphora resolution systems. A number of algorithms for identifying discourse-new mentions have therefore been proposed in the anaphora resolution literature, especially for definite descriptions. Vieira and Poesio use a wide range of hand-crafted heuristics[68] (these heuristics are discussed in some detail in Chapter 3). Bean and Riloff make use of syntactic heuristics, but also mine additional patterns for discourse-new description from corpus data [4]. Ng and Cardie develop a supervised algorithm, integrating a number of syntactic and lexical clues as features in a learning-based approach [42, 41]. Uryupina proposes a web-based algorithm for identifying discourse-new and unique NPs [62]. The approach helps overcome the data sparseness problem of [4]. by relying on Internet counts. Finally, Kabadjov provides an extensive evaluation of different features relevant for the discourse new detection task [29]. We will discuss the latter approach in Section 13.4 below. Anaphoricity detection algorithms have been also proposed for other languages, including Spanish [44] and Chinese [34]. We discuss below how these discourse-new detectors were integrated in anaphora resolution systems.

A corpus specific problem arose within the context of the recent CoNLL evaluation campaigns [50, 51]. The ONTONOTES dataset, as used for the CoNLL shared

tasks, only provides annotations for non-singleton coreference chains. Most participants have therefore applied some form of rule-based prefiltering to identify nominal expressions that are extremely unlikely to participate in any coreference relations, such as named entities of certain types (for example, QUANTITY) or non-referential pronouns (for example, "nobody"). To our knowledge, only very few approaches have been proposed to tackle the problem of ONTONOTES *mention detection* in a more principled way. Thus, Björkelund and Farkas have trained three classifiers to filter out non-anaphoric instances of *it*, *you* and *we* [7] . Kummerfeld et al. investigated various post- and pre-filtering heuristics for adapting their mention detection algorithm to the ONTONOTES English data in a semi-automatic way, reporting mixed results [35]. Finally, Uryupina and Moschitti propose a tree kernel-based algorithms for learning ONTONOTES mentions directly from the dara [64].

13.2.2.2 Integrating Anaphoricity and Referentiality Detectors into Coreference Resolution Systems

Early studies reported mixed results with respect to the effectiveness of discourse-new and referentiality detectors for full-scale coreference resolution. When such a detector is integrated as a preprocessing filter, some types of errors might propagate to drastically deteriorate the performance level of the coreference resolver: if a (referring or anaphoric) nominal expression is incorrectly filtered out at an early step, the system excludes it from the pool of candidate anaphors and can therefore never recuperate proposing a correct antecedent; on the contrary, if a (non-referring or discourse new) nominal expression is missed by the filter, the system might still correctly assign it to a singleton chain at a later stage by suggesting no suitable links. One of the most commonly used coreference resolution evaluation metric, the MUC scorer [69], is particularly sensitive to the former type of errors. With the introduction of more varied evaluation metrics (see Chapter 5 for details), however, the research on anaphoricty and referentiality detection received a considerable boost, leading to more advanced architectural solutions that, in turn, bring an improvement in the MUC score as well.

Ng and Cardie [41] propose a global optimization approach: they fine-tune learning parameters of their anaphoricity classifier to optimize the final performance level of the coreference resolver on the held-out data. Their experiments suggest that an anaphoricity pre-filter can improve the overall performance level of a coreference resolver, provided it is trained with an unbalanced precision-recall trade-off parameter. (See also Chapter 9.) A similar technique has been adopted by Uryupina and Moschitti, who use an oracle-based simulation experiment to limit the search space for several parameters to be optimized to reduce the amount of expensive re-training in global parameter optimization [64]. Kabadjov [29] advocates a post-filtering approach; this approach is discussed in detail in Section 13.4. Finally, Denis and Baldridge [18] propose a joint model for anaphoricity and antecedenthood detection. Their approach relies on integer linear programming (ILP) to find a solution

that is consistent with anaphoricity constraints. This approach is discussed in detail in Chapter 11.

It must be noted that all these algorithms improve the overall performance of coreference resolution, but at the same time require much more computational resources than the simplistic pipelined architecture. Thus, the global optimization solution [41] relies on multiple learning iterations and is therefore slow at the training stage. The ILP solution, on the contrary, does not require extra training time, but assumes a computationally expensive joint inference technique at the testing stage.

13.2.3 Event Anaphora and Abstract Object Anaphora

Finally, all systems have to deal with expressions that although strictly speaking anaphoric, are beyond the scope of the system, or the annotated corpus used to train and test it, or both. Recall the demonstrative pronoun *that* (M_6) in example (13.1). Although this mention is a context-depending anaphoric expression, its antecedent is not an entity introduced by another nominal mention, but an **abstract object** introduced by the clause *generals have the power*. Such anaphoric expressions are known as **references to abstract objects** [1] or cases of **discourse deixis** [71].

Interpreting references to abstract objects is very hard, as we even lack an agreed upon theory of what type of abstract objects there are. Asher [1] proposes an ontology including objects of increasing abstractness, from events to situations to propositions to facts, but there is no widespread agreement on how to discriminate between these different types of abstract objects. (See [48] for an empirical analysis of the types of objects referred to by demonstrative *this*, and [32] for an analysis of references to abstract objects via demonstrative nominals with so-called **shell nouns**.) As a result, until recently there weren't either corpora annotated with this type of references, or anaphora resolution algorithms for resolving them– most anaphora resolution systems only try to recognize such references so as not to attempt interpreting them. But there have been some recent developments in this regard.

Hand-coded algorithms for resolving pronominal references to abstract objects in the most general case have been proposed by Eckert and Strube [21] and Byron [11]. Kolhatkar [32] proposed statistical algorithms for resolving shell nouns, using the corpus she annotated. Much more progress has been made on a specific type of abstract reference: **event coreference resolution**. This progress has been motivated by the appearance of the ONTONOTES corpus (see Chapter 4), in which coreference links between anaphoric expressions (mainly pronouns) and events / situations are annotated. In the recent CoNLL evaluation campaigns (see Chapter 6 and [51, 50]) such links were included in the evaluation set, although given the complexity of the task, most participating systems opted for never proposing clause-level antecedents. A number of proposals were however published after the datasets were released, using rich syntactic representations to identify and resolve (pronominal) references to abstract events [33, 15].

13.2.4 Annotated Resources

The tasks of anaphoricity and referentiality detection have not received much attention from the computational linguistics community until recently. As a result, only very few manually annotated corpora support the development of supervised models for the tasks discussed in this Section by including annotations for coreference, anaphoricity and referentiality of the same documents. We briefly review here the options available–see Chapter 4 for a more extensive discussion of the available corpora.

Anaphoricity

Adopting a naive definition of discourse new mention as one that does not have any antecedent, we could straightforwardly induce the gold labels for anaphoricity from any coreference-annotated corpus: we mark the first mention in each chain as *discourse new*, whereas all the subsequent ones, if any, will be marked as *discourse old*. Such annotation scheme, however, only reflects a very simplistic definition of discourse novelty, covering only the mentions that Prince would classify as textually evoked descriptions. This situation is worsened by the fact that, as discussed in Chapter 4, in many of the most commonly used corpora singletons are not annotated: this is true, e.g., of MUC and ONTONOTES. Singletons are annotated in the ACE corpora, but only for the mentions referring to the restricted number of semantic types considered of interest. Clearly, such corpora do not provide very useful resources to train a discourse-new detector.

Three options are in principle available. First of all, a corpus in which all anaphoric relations are annotated could be used, including reference to abstract objects and associative anaphora. As far as we know there is only one medium-size corpus of this type, ARRAU [45] (see Chapter 4). The second option is to use a corpus in which although not all anaphoric relations are annotated, all mentions are marked with their information status, as done, e.g., in [43, 16, 39]. Finally, one could use a corpus in which only the information status of nominals is marked. All of these solutions of course require all referring mentions to be annotated.

Referentiality

Very few corpora label the referential status of noun phrases. The most commonly used datasets, MUC, ACE and ONTONOTES [25, 20, 50] do not include any annotation for non-referring NPs. According to the guidelines for these datasets, such NPs should be ignored by the coders and not treated as mentions to annotate. The problem is that one cannot assume that if an NP is not annotated it must be non-referential, since none of these corpora label all the referential expression: MUC and ONTONOTES do not mark singleton chains, whereas the ACE annotation is restricted to specific semantic types.

status	RST	Trains	PearStories
referential	62461 (86.73%)	14646 (86.15%)	3401 (84.85%)
non-referential: expletive	444 (0.61%)	853 (5.01%)	122 (3.04%)
non-referential: predicate	4387 (6.09%)	147 (0.86%)	84 (2.09%)
non-referential: coordination	2414 (3.35%)	235 (1.38%)	37 (0.92%)
non-referential: idiomatic	639 (0.88%)	149 (0.87%)	42 (1.04%)
non-referential: quantifier	1738 (2.41%)	818 (4.81%)	132 (3.29%)
non-referential: incomplete	2 (0%)	149 (0.87%)	36 (0.89%)

Table 13.1 Referentiality in the ARRAU corpus

The already mentioned ARRAU corpus however does include an extensive annotation of non-referential descriptions. According to the ARRAU guidelines, all nominal expressions, regardless of their referentiality, should be treated as mentions. Each mention is then labelled as referential or non-referential. Referential expressions are then further marked as discourse-new or discourse-old, and discourse-old should then be linked to their coreference chains. Non-referential mentions are then further subcategorized into expletives (13.4); predicates, including appositions, copulas and other types of predicative nominals (13.5); coordinations (13.6); parts of multiword or idiomatic expressions (13.7); quantifiers (13.8) and incomplete descriptions (13.9):

(13.4) But [it] doesn't take much to get burned.
(13.5) The new ad plan from Newsweek, [a unit of the Washington Post Co.], is [the second incentive plan the magazine has offered advertisers in three years].
(13.6) [Both Newsweek and U.S. News] have been gaining circulation in recent years..
(13.7) Apple II owners, for [example], had to use their television sets as screens and stored data on audiocassettes.
(13.8) They also said that vendors were delivering goods more quickly in October than they had for [each of the five previous months].
(13.9) what about [the uh]

Some statistics about referentiality for the different ARRAU domains are summarized in Table 13.1.

On average, around 13-15% of all the mentions in ARRAU are non-referential. This highlights the importance of the reference detection subtask for coreference resolution. Three ARRAU domains, however, exhibit different distributions of non-referential mentions. The news text from the WSJ portion of the Penn Treebank in the RST domain contain a large number of predicates and coordinations. Dialogue transcripts (Trains) and child fiction (PearStories), on the contrary, mostly contain very simple sentences with few nominal predicates. Expletive pronouns are very common in dialogues and fiction, but much more rare in news text. Finally, carefully edited RST documents contain virtually no incomplete mentions.

Other medium-to-large-scale corpora annotated with referentiality information include the ANCORA corpus for Spanish [57], the LiveMemories corpus for Italian [59] and the Tüba/DZ corpus of coreference in German.[2]

Discourse deixis

A number of corpora are annotated with references to abstract objects of different type. Pronominal references to abstract objects are annotated in the annotation of the TRAINS corpus produced by Byron [10], in the DAD corpus of Danish and Italian produced by Navarretta [40] and in the annotation of EuroParl by Dipper and Zinsmeister [19]. The corpus produced by Kolhatkar [32] contains annotated shell noun references. Among the general-purpose anaphora corpora, ARRAU and ANCORA are annotated for general reference to abstract objects, whereas in ONTONOTES, reference to events is annotated.

13.3 Detecting Non-Referentiality: Bergsma et al's algorithm

In this Section we discuss in more detail one of the best-known proposal of this type, the approach developed by Bergsma et al. [6, 5] that combines supervised learning with lexical features extracted from a large unlabeled dataset to create a robust and efficient system for detecting non-referential *it*. The approach relies on a binary classifier encoding syntactic and semantic properties of pronominal contexts. In what follows, we describe the methodology, two types of features used in the system, and evaluation results.

13.3.1 Methodology.

This algorithm models the non-referentiality detection task as a simple binary supervised classification problem. Each instance of *it* in the dataset is represented as a feature vector. The features encode various properties of the pronoun's context and are described in more details below. The classifier can be trained on any corpus annotated with referentiality. However, as we have seen in Section 13.2.4 above, such datasets are not common. To overcome the issue, referentiality labels can be induces from data annotated with coreferential links: if an instance of *it* is not linked to any other mention, one can assume that it is non-referential. Note that the same technique cannot be applied to other types of mentions: for example, referring nominal mentions can form singleton chains (consider mention "lowly soldiers" (M_2) in our

[2] http://www.sfs.uni-tuebingen.de/en/ascl/resources/corpora/tueba-dz.html

feature template	window size	examples
3, 4 and 5-grams containing `it`	-5,+5	it-is-able, it-is-able-to, it-is-able-to-maintain
token+position	-2,+5	is_{+1}, able_{+1}, …, maintain_{+4}
unigrams to the right	1,+20	is_{right}, able_{right}, to_{right}, maintain_{right}
unigrams to the left	-10,-1	$\emptyset$

Table 13.2 Detecting referentiality: templates for lexical features from [6] and their instanciations for (13.10A)

example (13.1)). A referential usage of *it*, however, should be either anaphoric or cataphoric.

Bergsma et al. report their evaluation results on two corpora: the BBN Pronoun Coreference Corpus [72] and the ItBank [5]. Both dataset contain newswire documents, however, they represent different domains corresponding to several text sources (Wall Street Journal, Science News and Slate).

Any off-the-shelf machine learning tool can be used for this problem, since it is formulated as a straightforward supervised binary classification task. In practice, Bergsma et al. report their results using the logistic regression packages from Weka and LibLinear.

13.3.2 Lexical Features.

Most traditional approaches to non-referential pronoun detection rely on syntactic heuristics to identify expletive usages of *it*. While syntactic patterns may provide important clues, they alone are not sufficient to discriminate between anaphoric and expletive pronouns. Consider the following example from [6]:

(13.10) A. [It] is able to maintain a stable price.
B. [It] is important to maintain a stable price.

Although these two sentences follow the same syntactic structure, in (13.10A), the pronoun refers to a specific object, whereas (13.10B) is a typical example of expletive *it*.

Following this observation, Bergsma et al. propose to focus on shallow context features. Provided sufficient training data are given to the model, it can be expected to learn the relevant patterns automatically without expensive manual syntactic feature engineering and, at the same time, making use of relevant lexical distinctions. To this end, each sentence containing *it* is first normalized: all the digits are replaced with 0, all the named entities are converted into a special "NE" token. The system then extracts lexical features encoding pronominal contexts, as summarized in Table 13.2.

filler type	string
3rd person singular pronoun	it/its
3rd person plural pronoun	they/them/their
any other pronoun	I/me/my/..
infrequent token	<UNK>
any other token	*

Table 13.3 Detecting referentiality: filler types for web-count features, as reported in [5].

13.3.3 Web Count Features.

A supervised model trained with the lexical features presented in Table 13.2 shows a promising performance level for such a simple approach. Thus, it can successfully induce the information that the presence of additional third person neuter pronouns ("it" or "its") is a good indicator for referentiality. On the contrary, complementizers ("that", "to") often occur within expletive contexts.

However, on a conventional size corpus, the model suffers from the data sparsity problem, making it impossible to fully represent the relevant lexical information. To overcome the issue, Bergsma et al. rely on the information extracted from web counts. From the manually annotated data, they extract contexts containing the pronoun "it" and use them to generate templates, removing the pronoun itself. At the next stage, they query the Google N-gram corpus to obtain counts for different words filling the empty slot. This approach relies on the following intuition: in referential contexts, non-pronominal words can appear relatively often (for example, "China is able to maintain", generated from (13.10A) could be found multiple times in a large enough corpus); expletive contexts, on the contrary, generate patterns that are characteristic for non-referential pronouns and cannot be filled with other words ("China is important to maintain" would be rare even in a very large dataset).

Following a normalization step (stemming, applying heuristics for irregular verbs, replacing digits and named entities with special auxiliary tokens), Bergsma et al. extract web counts for different types of fillers, as summarized in Table 13.3. They encode obtained web counts as features, with each individual feature corresponding to the filler's type, the template length (4 or 5) and the position of the filler in the pattern.

In [6], a modification of this feature subset is proposed. The new features rely on a restricted set of patterns (only "it", "they" and "them" fillers and only 4-token templates are considered) and a more aggressive normalization strategy. This results in a minor performance drop, but at the same time allows for a very efficient implementation that can be stored in the memory together with all the models and thus allows for fast processing of large amounts of data.

features	BBN-test	WSJ-2	ItBank
Majority Class	72.5	74.9	67.7
Lexical Features	82.9	82.5	78.7
Web Counts	83.3	85.6	83.1
Lexical+Counts (NADA)	86.0	86.2	85.1

Table 13.4 Detecting referentiality: the system's accuracy (%) on different test sets, as reported in [6].

13.3.4 Evaluation.

Table 13.4 summarizes the evaluation results reported by Bergsma et al. for different parts of their system [6]. Note that the reported figures correspond to several test sets, with the train set remaining the same (roughly a half of the BBN corpus). Of the three test sets, two come from the same domain as BBN-train (BBN-test and WSJ-2), whereas ItBank contains out-of-domain data. The results show several important trends. First, only 67-75% of "it" instances are referential. This highlights the importance of the task for accurate pronoun resolution and text understanding in general. Second, web counts outperform more traditional lexical features. Although web counts correspond to more shallow patterns and smaller contexts, they are extracted from a much larger text collection and thus suffer less from the data sparsity issue. This is especially crucial for the out-of-domain setting, where web counts outperform lexical features by almost 5%. Most importantly, the two groups of features can be combined effectively to further boost the performance. This combination is used by NADA – publicly available toolkit for detecting non-referential "it" instances.[3]

13.4 Discourse-new Detection

Virtually every modern anaphoric resolver includes some ability to detect discourse-new mentions, but relatively few studies have focused exclusively on the issue. In this section we describe in detail one example of in-depth investigation of discourse-new detection and its use in anaphora resolution, the analysis of discourse-new detection initiated by Poesio *et al.* [47] and continued in Kabadjov's doctoral thesis [29].

[3] `https://code.google.com/p/nada-nonref-pronoun-detector/`

13.4.1 Features for Discourse New Detection

The objective of Poesio *et al.* was to assess the results of integrating a discourse-new detector with a definite description resolver, specifically the DD resolution module of GUITAR [46].[4] To begin with, Poesio *et al.* [47] analyzed the literature on discourse-new detection define their feature space as follows. Every definite description from the annotated corpus produces a training example in the form of $(\mathbf{x}_j, c)$. The target c is binary, the DD has an annotated antecedent or not (i.e., discourse-new). The input features $(\mathbf{x})$ attempt to capture various types of knowledge and are discussed next.

1. **Direct Anaphora.** A single feature, which is produced by running the *direct anaphora* algorithm proposed by Vieira and Poesio [68] and recording the distance in terms of utterances between the antecedent put forward by the algorithm and the anaphor (i.e., the definite description being resolved). The possible values for this features are -1, 0, 1, 2, ..., which respectively mean no antecedent was proposed, the antecedent is within the same utterance, one utterance apart, two utterances apart, and so on.

Predicative. Two features to detect *predicative noun phrases*:

2. **Apposition.** Value 1 if the DD appears in appositive position, 0 otherwise.
3. **Copular.** Value 0 if not in a copular sentence, 1 if the DD appears on the left-hand side of the copula, and 2 if it appears on the right-hand side.

Proper names. Two boolean features to capture *proper names*:

4. **C-head.** Value 1 if the DD head is capitalised, 0 otherwise.
5. **C-premod.** Value 1 if one of the premodifiers is capitalised, 0 otherwise.

Functionality. As for *funtional* definite descriptions, there are several features meant to approximate "definite probability" by making use of Internet counts. First, they compute Internet counts using Google's API for the following lexical level variations of each DD: "Det Y", "Det H" and "Det A", where `Y` is the phrase left after removing the DD determiner, `H` is the syntactic head of the DD, `A` is the first adjective premodifier (if any) and `Det` is either "the", "a" or "an". And then the following ratios are computed:

6. $\frac{\#\text{“}theH\text{”}}{\#\text{“}aH\text{”}}$
7. $\frac{\#\text{“}theH\text{”}}{\#H}$
8. $\frac{\#\text{“}theY\text{”}}{\#\text{“}aY\text{”}}$
9. $\frac{\#\text{“}theY\text{”}}{\#Y}$
10. $\frac{\#\text{“}theA\text{”}}{\#\text{“}aA\text{”}}$
11. $\frac{\#\text{“}theA\text{”}}{\#A}$
12. **Superlative.** Value 1 if one of the premodifiers is a superlative, 0 otherwise.
13. **Establishing Relative.** A single feature, whose value is 0 if the DD is not postmodified, 1 if postmodified by a prepositional phrase, 2 if postmodified by a relative clause, and 3 otherwise (i.e., some other type of postmodification).

[4] GUITAR's DD resolution module is is an implementation of the Vieira / Poesio algorithm, see Chapter 3.

Position. Three features to capture the *position* of the definite description within the text:

14. **Title.** Value 1 if the DD appears in the title (if markup available), 0 otherwise.
15. **FirstPar.** Value 1 if the DD appears in the first paragraph of the document (if markup available), 0 otherwise.
16. **FirstSent.** Value 1 if the DD appears in the first sentence of the document, 0 otherwise.

Surface. A few features to capture *surface features* of the definite description:

17. **DDHasAttributes.** Value 1 if the set of attributes (i.e., premodifiers) is not empty, 0 otherwise.
18. **IsEmbedded.** Value 1 if the definite description is an embedded noun phrase (possibly postmodifying another NP).
19. **NumberOfWords.** The number of words composing the definite description.

V&P. Three features from the original *Vieira and Poesio's algorithm*:[5]

20. **DDAttributesIsSubset.** Value 1 if the set of attributes of DD is a subset of the set of attributes of the antecedent, 0 otherwise.
21. **AnteSubsequentMention.** Value 1 if the size of the equivalence class of the antecedent is greater than 1, 0 otherwise.
22. **AnteIdentical.** Value 1 if the proposed antecedent and the anaphor are identical definite descriptions, 0 otherwise.

Antecedent. Several features representing knowledge about the *proposed antecedent*:[6]

23. **AnteHasAttributes.** Value 1 if the set of attributes of the proposed antecedent is not empty, 0 otherwise.
24. **AnteIsEmbedded.** Value 1 if the proposed antecedent is an embedded noun phrase.
25. **AnteType.** This is a number from 0 to 19 uniquely identifying the type of NP that is the proposed antecedent (e.g., the-np, the-pn, pn).

13.4.2 Integrating a discourse-new detector in a definite description resolver

As discussed in Section 13.2.2.2, four main approaches to integrating discourse-new detection into coreference resolution have been pursued in the literature:

1. Do not train a separate classifier for discourse-new detection, but allow the coreference classifier to use features such as those discussed in the previous Section directly. This is arguably the approach most commonly adopted.
2. Train a separate discourse-new detector, and use it as a *pre- resolutionfilter* removing candidate mentions prior to coreference resolution. This is the approach followed by, for instance, Ng and Cardie [41]. (See also the discussion of their approach in Chapter 9.)

[5] In the case where no antecedent has been proposed, these features assume a value of -1.

[6] In the case where no antecedent has been proposed, these features assume a value of -1.

3. Use the discourse-new detector as a *post resolution* filter, either allowing an antecedent proposed by the coreference resolver to go through, or else precluding an anaphoric link. This is the approach followed by Kabadjov [29]; we discuss it in detail this Section.
4. Use Integer Linear Programming (ILP) to integrate two separate classifiers, a coreference resolver and a discourse-new detector, as proposed by Denis and Baldridge [18]. This approach is discussed in Chapter 11.

Kabadjov [29] carried out a number of experiments using the features discussed above to train a discourse-new detector, and integrating the detector as a post-resolution filter for the the definite description resolver of GUITAR.

In [29] experiments with four machine learning methods are discussed: decision trees, maximum entropy (ME), support vector machines (SVMs) and neural networks. Here we focus only on their work with ME and SVM classifiers, since these were used in both of their main experiments (discussed below). For the SVM experiments, Kabadjov used the implementation of the SVM technique in the publicly available LIBSVM library [13] with radial basis function (RBF) kernel $K(x,y) = e^{-\gamma||x-y||^2}$ and the optimal parameters C and γ obtained by cross-validation within each fold. For the ME experiments, Kabadjov used the openNLP MAXENT package [3] which is a library for training and using maximum entropy models. The implemented training algorithm in the library is the *Generalised Iterative Scaling* algorithm using 100 iterations. ME models output a probability for a given class, which means a threshold must be used to produce an actual classification. Kabadjov [29] found an optimal threshold of 0.74 for the discourse-new class based on his cross validation experiment.

Kabadjov [29] ran two experiments: a ten fold cross validation experiment on hand-parsed data and an experiment on automatically parsed data. The GNOME and VPC corpora were used as datasets in both experiments. We summarise both below.

Experiment 1: Processing Hand-parsed Data.

In order to evaluate separately the performance of the system at detecting Non-Anaphoricity (NA), Coreference Resolution (COREF), and overall, Kabadjov introduced separate P/R metrics, defined as follows. The definitions of P/R used for evaluating performance at NA are:

$$P_{NA} = \frac{NA_{corr}}{NA_{sys}} \tag{13.11}$$

$$R_{NA} = \frac{NA_{corr}}{NA} \tag{13.12}$$

where NA_{corr} is the number of markables correctly classified as non-anaphoric, NA_{sys} is the total number of markables classified as non-anaphoric and NA is the total number of targeted non-anaphoric markables.

For coreference resolution, he introduces the following versions of precision and recall:

$$P_{COREF} = \frac{COREF_{corr}}{COREF_{sys}} \tag{13.13}$$

$$R_{COREF} = \frac{COREF_{corr}}{COREF} \tag{13.14}$$

where $COREF_{corr}$ is the number of markables resolved correctly to their antecedent, $COREF_{sys}$ is the number of markables for which an antecedent was proposed by the system and COREF is the total number of markables that have an annotated antecedent.

Finally, Kabadjov introduces a combined measure of performance taking into account both non-anaphoricity classification and coreference resolution:

$$P = \frac{NA_{corr} + COREF_{corr}}{NA_{sys} + COREF_{sys}} \tag{13.15}$$

$$R = \frac{NA_{corr} + COREF_{corr}}{NA + COREF} \tag{13.16}$$

where NA_{corr}, $COREF_{corr}$, NA_{sys}, $COREF_{sys}$, NA and COREF are as defined above.

On discourse-new classification Kabadjov's ME and SVM classifiers attained $F_{NA} = 89\%$ and $F_{NA} = 90.2\%$, respectively, compared to $F_{NA} = 78.9\%$ obtained by the baseline of assigning all definite descriptions to the discourse-new class, and an upper bound of $F_{NA} = 100\%$.[7].

On definite description resolution Kabadjov's ME- and SVM-based modules achieved $F_{COREF} = 68.6\%$ and $F_{COREF} = 68.8\%$ respectively, whereas Vieira and Poesio's direct anaphora algorithm (henceforth, baseline$_1$) achieved $F_{COREF} = 67.9\%$ and a simple same-head match algorithm, which proposes as antecedent the last markable (if any) with the same head occurring before the definite description in the text (henceforth, baseline$_2$) achieved $F_{COREF} = 62.7\%$.The upper bound obtained if perfect DN classification (i.e., 100%) was available was $F_{COREF} = 74.2\%$. Both ME- and SVM-based modules were statistically significantly superior to both baselines according to the *t* and *sign* tests.

Finally, the overall performance of Kabadjov's ME- and SVM-based modules was $F = 83.6\%$ and $F = 83.5\%$, respectively, in the context of the same baselines as in the previous case, $F = 82.1\%$ and $F = 75.22\%$, for baseline$_1$ and baseline$_2$, respectively, and a ceiling of $F = 91.8\%$.[8]

Kabadjov [29] carried out a sample complexity analysis to estimate the minimum size of training set for successful learning to happen. Based on the lack of performance peak over a gradual increase of performance by increasing the data set, he concluded that more data was needed to reach optimum performance, though, based

[7] However, note that a ceiling of 100% may be a bit too high based on earlier work by Poesio and Vieira [49] where inter-annotator agreement on the task was estimated at $K = 0.76$ (kappa value).

[8] The baselines and ceiling scores were computed with some additional assumptions such as considering discourse-new those definite descriptions for which no antecedent was proposed by the original resolution algorithm.

on their experiments he noted that having a discourse-new detector even if not maximally accurate is better than having no discourse-new detection in place.

Additionally, Kabadjov also performed a feature analysis to gain an insight into what features are more important than others. For that purpose he analysed the induced decision trees. He found out that the most important feature was the *direct anaphora* feature, since it consistently appeared at the root of all the decision trees, followed by the length of the definite description (number of words) and whether there is a capitalised premodifier (cPremod). Kabadjov also noted that the two most common leaf nodes were the *AnteCat* feature capturing the type of the antecedent (e.g., proper name, definite description, pronoun, etc.) and the *Relative* feature, which captures the type of post-modification.

Experiment 2: Processing Automatically Parsed Data.

In addition to the 10-cross-validation experiment on hand-parsed data, Kabadjov also ran an experiment on fully automatically processed texts. The purpose of this second experiment was to compare learned classifiers trained on 93% of the data set from the 10-X-Validation experiment[9] (i.e., 93% of GNOME and VPC combined) and tested on a different data set (the CAST corpus).

In order to carry out this experiment Kabadjov parsed the CAST corpus with Charniak's parser [14] and then automatically aligned the noun phrases (NPs) with the markables of human annotation. He also modified the way the *sign* test is applied by considering as test items the definite descriptions, for which the version of GUITAR without the DN classifier proposes an antecedent, and then within this set DD's positively affected by the DN classifier count as pluses, DD's negatively affected as minuses and DD's on which DN has made no difference are ignored.

The ME- and SVM-based modules of the system achieved resolution performance of $F_{COREF} = 58.9\%$ and $F_{COREF} = 59.3\%$, respectively, in the context of $F_{COREF} = 58.5\%$ as baseline[10] and $F_{COREF} = 65.4\%$ as ceiling[11]. In this second experiment only their SVM-based module achieved statistically significant improvement over the baseline.

In the light of the performance upper bounds presented above, Poesio *et al.*'s and Kabadjov's experimental results suggest that there is clearly still much room for improvement. On one hand Kabadjov claims that extending their experiments on a larger data set will produce more accurate DN classifiers which will in turn translate into improved definite description resolution. On the other hand extending the data representation model to include more features that are able to capture more fine-grained aspects and subtleties of the discourse-new phenomenon as well as more accurate pre-processing for computing feature values would naturally raise the overall performance.

---

[9] The remaining 7% were set aside for validation and parameter tuning.

[10] The baseline here is baseline$_1$ from the 10-X-Validation experiment (i.e., Vieira and Poesio's direct anaphora algorithm).

[11] Assuming perfect DN classification.

13.5 Conclusions

In this chapter we surveyed work on non-reference and non-anaphoricity detection, and discussed in detail the proposal for expletive detection by Bergsma *et al.* [6, 5] and by Poesio *et al.* and Kabadjov on discourse-new recognition [47, 29].

Acknowledgments

This work was supported in part by the LIMOSINE project (Uryupina, Poesio), in part by the SENSEI project (Kabadjov, Poesio).

References

[1] Asher, N.: Reference to Abstract Objects in English. D. Reidel, Dordrecht (1993)
[2] Asher, N., Wada, H.: BUILDRS: An implementation of DR theory and of LFG. In: Proc. of COLING-86, pp. 540–545. Bonn, FRG (1986)
[3] Baldridge, J., Morton, T.: The openNLP MAXENT package. Software available at http://maxent.sourceforge.net/
[4] Bean, D.L., Riloff, E.: Corpus-based identification of non-anaphoric noun phrases. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (1999)
[5] Bergsma, S., Lin, D., Goebel, R.: Distributional identification of non-referential pronouns. Proc. of ACL-08: HLT pp. 10–18 (2008)
[6] Bergsma, S., Yarowsky, D.: NADA: A robust system for non-referential pronoun detection. In: Proc. DAARC, pp. 12–23. Faro, Portugal (2011)
[7] Björkelund, A., Farkas, R.: Data-driven multilingual coreference resolution using resolver stacking. In: Joint Conference on EMNLP and CoNLL - Shared Task, pp. 49–55. Association for Computational Linguistics, Jeju Island, Korea (2012). URL http://www.aclweb.org/anthology/W12-4503
[8] Bos, J.: Wide-coverage semantic analysis with BOXER. In: J. Bos, R. Delmonte (eds.) Semantics in Text Processing. Proceedings of STEP, pp. 277–286 (2008)
[9] Boyd, A., Gegg-Harrison, W., Byron, D.: Identifying nonreferential *it*: A machine learning approach incorporating linguistically motivated patterns. In: Proceedingd of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing,, pp. 40–47 (2005)
[10] Byron, D.: Resolving pronominal reference to abstract entities. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (2002)

[11] Byron, D.K.: Resolving pronominal reference to abstract entities. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02), pp. 80–87 (2002)
[12] Byron, D.K., Gegg-Harrison, W.: Eliminating non-referring noun phrases from coreference resolution. In: In Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2004), pp. 21–26 (2004)
[13] Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
[14] Charniak, E.: A maximum-entropy-inspired parser. In: Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) (2000)
[15] Chen, B., Su, J., Pan, S.J., Tan, C.L.: A twin-candidate based approach for event pronoun resolution using composite kernel. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING) (2013)
[16] Collovini, S., Vieira, R.: Learning discourse-new references in portuguese text. In: Proc. of IFIP 19th World Computer Congress, pp. 267–276. Springer, Santiago, Chile (2006)
[17] Daelemans, W.: TiMBL: Tilburg University Memory Based Learner version 2 Reference Guide. Tech. Rep. ILK99-01, Tilburg University (1999)
[18] Denis, P., Baldridge, J.: Joint determination of anaphoricity and coreference resolution using integer programming. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, pp. 236–243 (2007)
[19] Dipper, S., Zinsmeister, H.: Annotating abstract anaphora. Language Resources and Evaluation **46**(1), 37–52 (2012)
[20] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassell, S., Weischedel, R.: The automatic content extraction (ACE) program–tasks, data, and evaluation. In: Proceedings of the Language Resources and Evaluation Conference (2004)
[21] Eckert, M., Strube, M.: Dialogue acts, synchronising units and anaphora resolution. Journal of Semantics (2001)
[22] Evans, R.: Applying machine learning toward an automatic classification of *it*. Literary and Linguistic Computing **16**(1), 45–57 (2001)
[23] Fraurud, K.: Definiteness and the processing of NPs in natural discourse. Journal of Semantics **7**, 395–433 (1990)
[24] Hawkins, J.A.: Definiteness and Indefiniteness. Croom Helm, London (1978)
[25] Hirschman, L.: MUC-7 coreference task definition, version 3.0. In: N. Chinchor (ed.) Proceedings of the 7th Message Understanding Conference. NIST (1998). Available online at http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html
[26] Hirst, G.: Anaphora in Natural Language Understanding: A Survey. Springer (1981)
[27] Hobbs, J.: Resolving pronoun references. Lingua **44**(311), 339–352 (1978)
[28] Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R.: OntoNotes: The 90% solution. In: Proceedings of HLT/NAACL (2006)

[29] Kabadjov, M.A.: A comprehensive evaluation of anaphora resolution and discourse-new recognition. Ph.D. thesis, Department of Computer Science, University of Essex (2007)
[30] Kamp, H., Reyle, U.: From Discourse to Logic. D. Reidel, Dordrecht (1993)
[31] Karttunen, L.: Discourse referents. In: J. McKawley (ed.) Sytax and Semantics, vol. 7, pp. 361–385. Academic Press (1976)
[32] Kolhatkar, V.: Resolving shell nouns. Ph.D. thesis, University of Toronto (2014)
[33] Kong, F., Zhou, G.: Improve tree kernel-based event pronoun resolution with competitive information. In: Proceedings of IJCAI (2011)
[34] Kong, F., Zhu, Q., Zhou, G.: Anaphoricity determination for coreference resolution in english and chinese languages. Journal of Computer Research and Development **49**(5), 1072 (2012)
[35] Kummerfeld, J.K., Bansal, M., Burkett, D., Klein, D.: Mention detection: Heuristics for the OntoNotes annotations. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pp. 102–106. Association for Computational Linguistics, Portland, Oregon, USA (2011). URL `http://www.aclweb.org/anthology/W11-1916`
[36] Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4), 535–562 (1994)
[37] Loebner, S.: Natural language and generalised quantifier theory. In: P. Gärdenfors (ed.) Generalized Quantifiers, pp. 93–108. D. Reidel, Dordrecht, The Netherlands (1987)
[38] Mitkov, R.: Anaphora Resolution. Longman (2002)
[39] Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J.: Ancor_centre, a large free spoken french coreference corpus. In: Proc. of LREC (2014)
[40] Navarretta, C.: Pronominal types and abstract reference in the danish and italian dad corpora. In: C. Johansson (ed.) Proc. of the Second Workshop on Anaphora Resolution (WAR II), pp. 63–71 (2008)
[41] Ng, V.: Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 152–159 (2004)
[42] Ng, V., Cardie, C.: Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In: Proceedings of the 19th International Conference on Computational Linguistics, pp. 730–736 (2002)
[43] Nissim, M., Dingare, S., Carletta, J., Steedman, M.: An annotation scheme for information status in dialogue. In: Proc. of LREC (2004)
[44] Palomar, M., Muñoz, R.: Definite descriptions in an information extraction system. In: M. Monard, J.a. Sichman (eds.) Advances in Artificial Intelligence, *Lecture Notes in Computer Science*, vol. 1952, pp. 320–328. Springer Berlin Heidelberg (2000)
[45] Poesio, M., Artstein, R.: Anaphoric annotation in the arrau corpus. In: Proc. of LREC. Marrakesh (2008)

[46] Poesio, M., Kabadjov, M.A.: A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal (2004)
[47] Poesio, M., Kabadjov, M.A., Vieira, R., Goulart, R., Uryupina, O.: Do discourse-new detectors help definite description resolution? In: Proceedings of the International Workshop on Computational Semantics (IWCS) (2005)
[48] Poesio, M., Modjeska, N.N.: Focus, activation, and this-noun phrases: An empirical study. In: A. Branco, R. McEnery, R. Mitkov (eds.) Anaphora Processing, pp. 429–442. John Benjamins (2005)
[49] Poesio, M., Renata, V.: A corpus-based investigation of definite description use. Computational Linguistics **24**(2), 183–216 (1998)
[50] Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL'12). Jeju, Korea (2012)
[51] Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011). Portland, Oregon (2011)
[52] Prince, E.F.: Toward a taxonomy of given-new information. In: P. Cole (ed.) Radical Pragmatics, pp. 295–325. Academic Press, New York (1981)
[53] Prince, E.F.: The ZPG letter: subjects, definiteness and information status. In: S. Thompson, W. Mann (eds.) Discourse description: diverse analyses of a fund-raising text, pp. 295–325. John Benjamins (1992)
[54] Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A Comprehensive Grammar of the English Language. Longman, Harlow, UK (1985)
[55] Recasens, M.: Coreferència: Teoria, anotació, resolució i avaluació. Ph.D. thesis, Universitat de Barcelona (2010)
[56] Recasens, M., de Marneffe, M.C., Potts, C.: The life and death of discourse entities: Identifying singleton mentions. In: HLT-NAACL, pp. 627–633 (2013)
[57] Recasens, M., Martí, M.A.: Ancora-co: Coreferentially annotated corpora for spanish and catalan. Language Resources and Evaluation (2009)
[58] Roberts, C.: Modal Subordination, Anaphora and Distributivity. Garland, New York (1990)
[59] Rodriguez, K.J., Delogu, F., Versley, Y., Stemle, E., Poesio, M.: Anaphoric annotation of wikipedia and blogs in the live memories corpus. In: Proc. LREC (poster) (2010)
[60] Sinclair, J. (ed.): Collins COBUILD English Grammar. Harper Collins, London (1995)
[61] Swan, M.: Practical English Usage. Oxford University Press (1995)
[62] Uryupina, O.: High-precision identification of discourse-new and unique noun phrases. In: Proceedings of the ACL'03 Student Workshop, pp. 80–86 (2003)
[63] Uryupina, O.: Detecting anaphoricity and antecedenthood for coreference resolution. Processamento del Lenguaje Natural **42** (2009)

[64] Uryupina, O., Moschitti, A.: Multilingual mention detection for coreference resolution. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP'13) (2013)
[65] Vallduvi, E.: Information packaging: a survey. Research Paper RP-44, University of Edinburgh, HCRC (1993)
[66] Versley, Y., Moschitti, A., Poesio, M., Yang, X.: Coreference systems based on kernels methods. In: Proceedings of the International Conference on Computational Linguistics (COLING), pp. 961–968 (2008)
[67] Vieira, R.: Definite description resolution in unrestricted texts. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh (1998)
[68] Vieira, R., Poesio, M.: An empirically-based system for processing definite descriptions. Computational Linguistics **26**(4), 539–593 (2000)
[69] Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 45–52 (1995)
[70] Ward, G., Birner, B.J.: Information structure. In: L.R. Horn, G. Ward (eds.) Handbook of Pragmatics, pp. 153–174. Oxford: Basil Blackwell (2004)
[71] Webber, B.L.: Structure and ostension in the interpretation of discourse deixis. Language and Cognitive Processes **6**(2), 107–135 (1991)
[72] Weischedel, R., Brunstein, A.: Bbn pronoun coreference and entity type corpus. LDC2005T33 (2005)